

## Data Products – From Design to Build, to Publishing and Consumption

Mike Ferguson  
Managing Director  
Intelligent Business Strategies

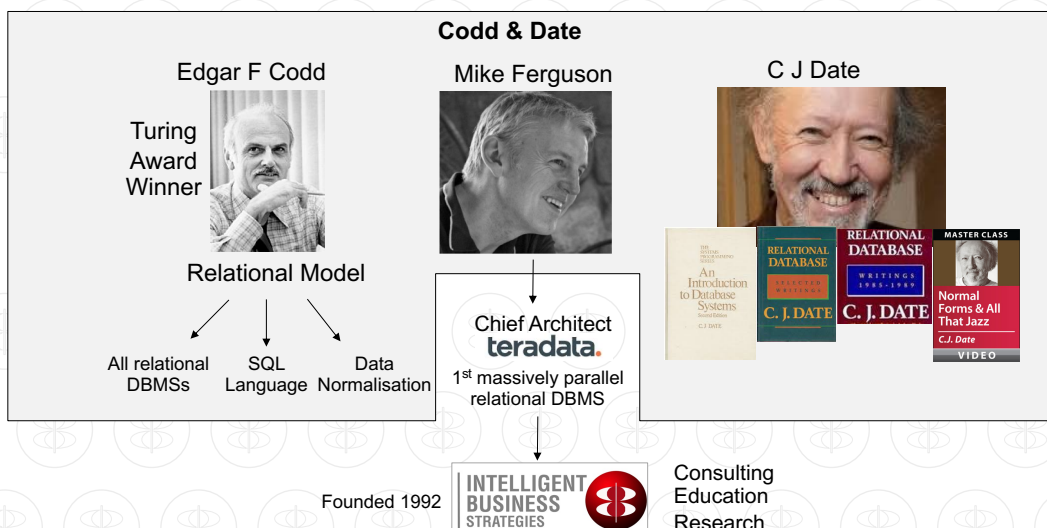


Adept Events Data Warehousing & BI Summit  
March 2024

 @mikeferguson1



## Who Is Mike Ferguson? – A World Leader In Data Management & Analytics





## Mike Ferguson Is Europe's Leading Industry Analyst / Consultant In Data Management & Analytics And Conference Chairman Of Big Data LDN

**BIG DATA LDN.**



Big Data LDN is the largest data & analytics conference in Europe

- 20000 delegates
- 180 vendors
- 15 theatres
- 300+ speakers

It is 6 x size of Gartner's D&A conference



Copyright © Intelligent Business Strategies 1992-2024



## About Intelligent Business Strategies

- A UK-based independent IT analyst and consulting firm founded 1992 specialising in data management & analytics
- Mike Ferguson is an independent IT Industry Analyst and consultant, Conference Chairman of Big Data LDN, and a member of the EDM Council CDMC Executive Advisory Board
- Our customers are Fortune 1000 companies, software vendors, venture capitalists, consulting firms and universities
- Three main lines of business in the areas of **Data Management, Analytics (BI / ML / AI) and Intelligent Business**

**Consultancy**

**Customers**

- CDO & Project Advisory
- Data & Analytics Strategy
- Data Architecture
- Data Governance
- Technology Selection
- Data & Analytics Reviews

**Vendors**

- Product & Go to Market Strategy
- Product Positioning
- Marketing Support
- Speaking at vendor events
- White Papers
- Webinars

**Venture Capitalists**

- Due-diligence, Asset advisory

**Education**

**Courses**

- Data Strategy
- Modern Data Architecture
- How to Govern a Distributed Data Estate
- Practical Guidelines for Implementing a Data Mesh
- Data Catalogs
- DW Modernisation
- DW Migration to the Cloud
- Embedded Analytics, Intelligent Apps & AI Automation
- Public classes (anyone)
- On-site classes (single client)
- On-line (public & on-sites)

We train customers, vendor and SIs

**Research**

**D&A Technology Research**

- Data Catalogues
- Data Fabric
- Data Governance
- DBMS Software
- Cloud Data & Analytics
- AI-Driven Automation
- Data science workbenches
- Generative AI

**Market Research**

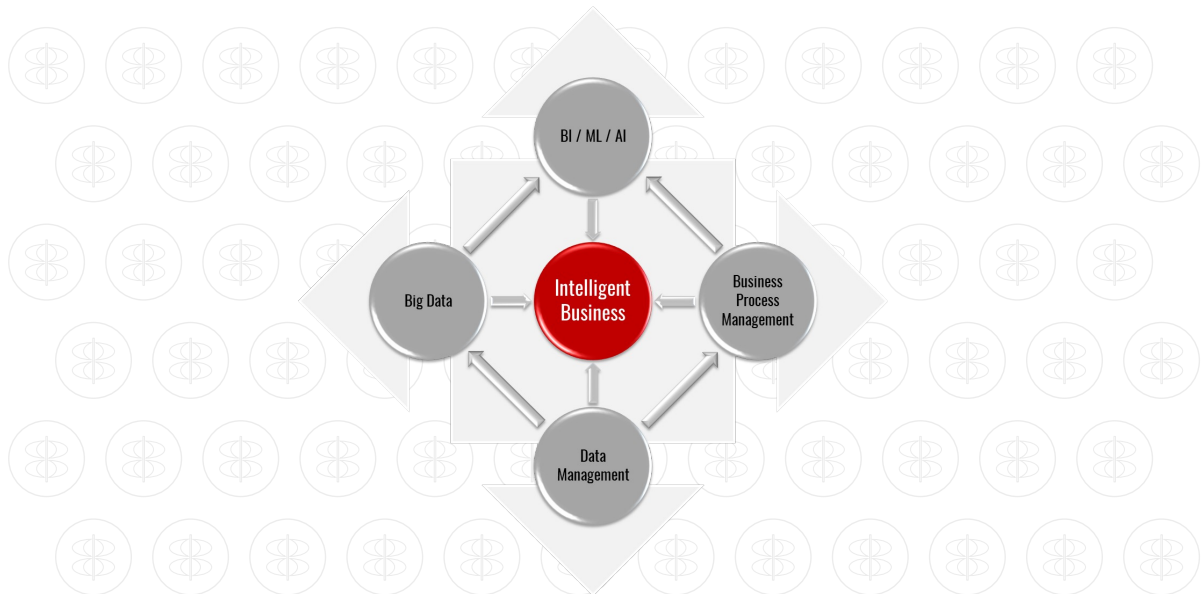
- 4th Industrial Revolution Survey

[www.intelligentbusiness.biz](http://www.intelligentbusiness.biz)

Copyright © Intelligent Business Strategies 1992-2024



## Our Focus Is Helping Companies Building The Intelligent Business



Copyright © Intelligent Business Strategies 1992-2024

5



## Topics

- What are data products?
- What makes creating data products different from other approaches to creating data for use analytical workloads?
  - How to design semantically linked data products to enable rapid consumption and use of data to produce new insights
  - Quick start mechanisms to speed up data product design
  - Defining common business data names for data products in a business glossary
  - Data modelling techniques for data products
  - Discovering data needed to build data products using a data catalog
  - Developing DataOps pipelines to engineer the data needed using data fabric
- Publishing data products – the role of the data marketplace
- Governing access to and use of data products across the enterprise
- Consuming and assembling data products for use in multiple analytical workloads
- Technologies and skills needed

Copyright © Intelligent Business Strategies 1992-2024

6

**Executive Expectations Of Data And AI Are Huge But Maximum Return On Investment Will Not Happen Unless There Is A Solid Foundation Of High-Quality, Reusable Data**

**DATA & AI**

Disruption Transformation

Speed & Agility Competitive Advantage

The Data and AI Driven Enterprise

High Quality & Governed Data Foundation

7

**Many Companies Don't Have A Data Foundation – What They Have Is Multiple Siloed Analytical Systems With Many Different Data Integration Tools And Multiple Copies Of Data**

Silo Silo Silo Silo Silo

Analytical models/tools/apps

Graph analysis tools

Notebooks, Mining tools ML models

BI tools

Transaction & Analytical systems

Graph Analysis

Data Science (any data)

DW & marts

MDM

NoSQL graph DB

Spark Cloud storage

EDW mart

CRM ERP SCM

CRM ERP SCM

Data integration tools

Data integration tools

Data integration tools

Data integration tools

Streaming data

Multi-structured & structured data

Multi-structured data

Structured data

Structured data

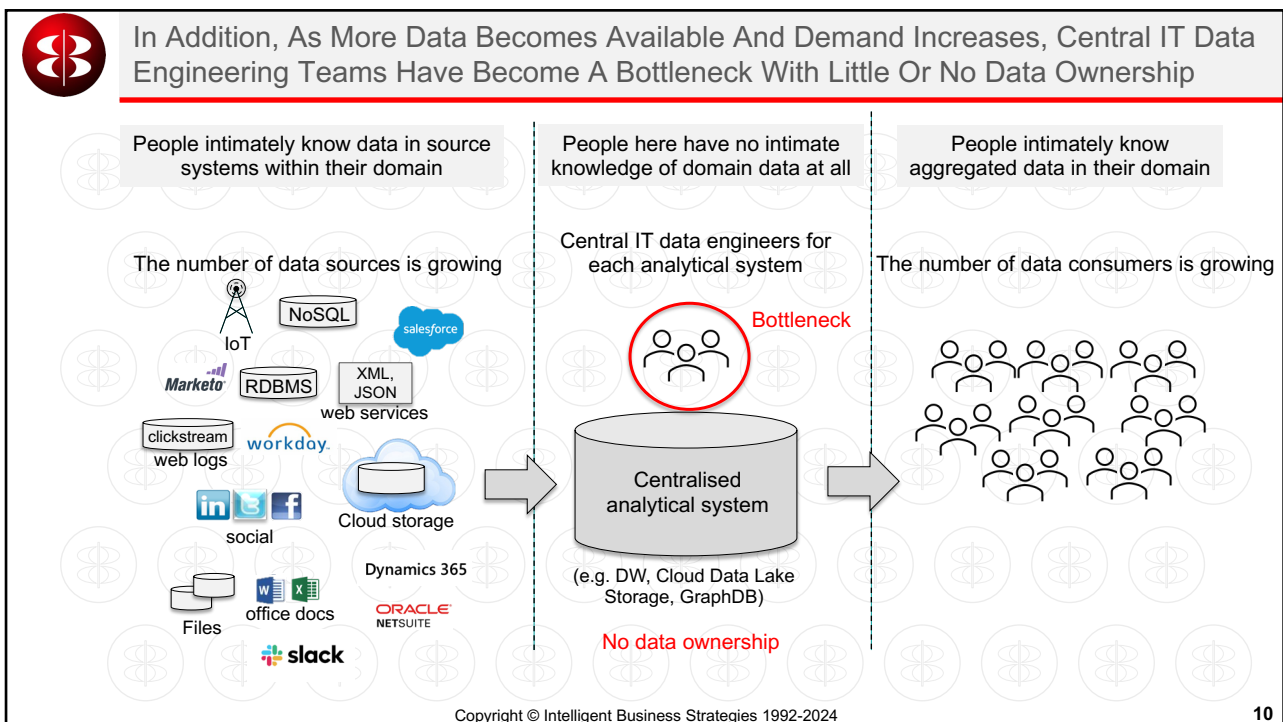
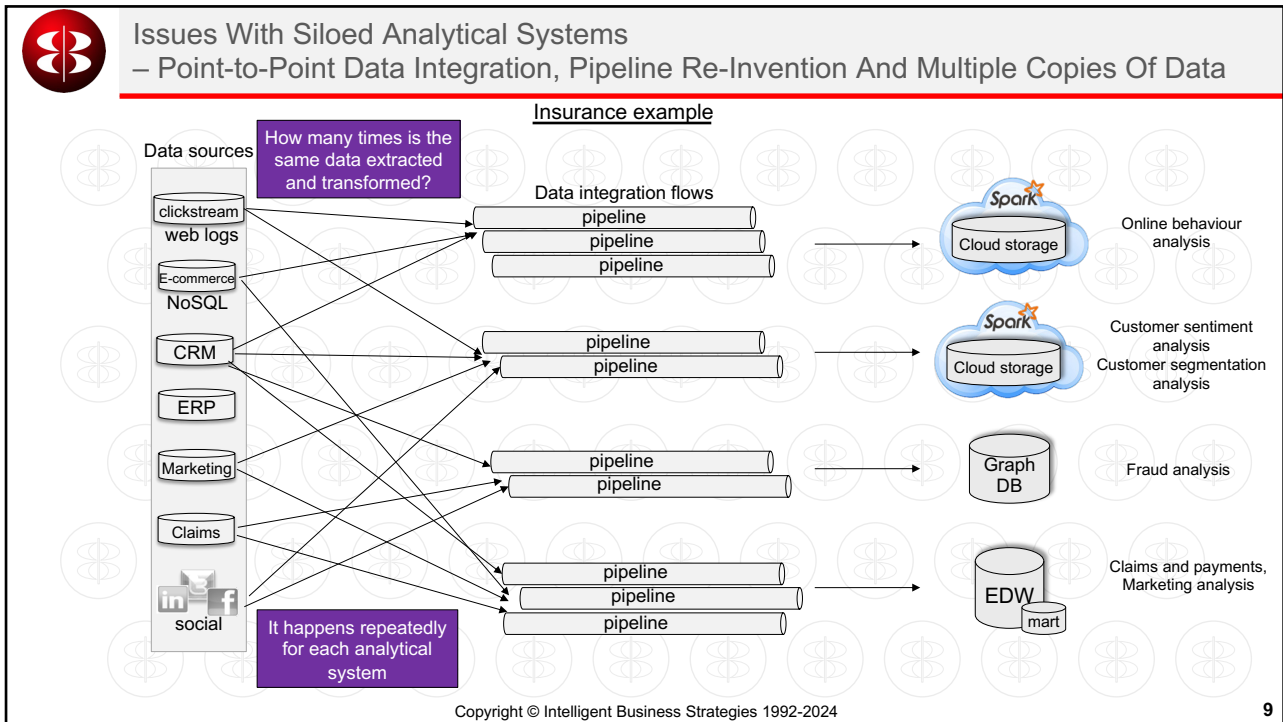
How many tools, scripts and programs are in use to clean/integrate data?

Unlikely that metadata is shared across tools

Copyright © Intelligent Business Strategies 1992-2024

8

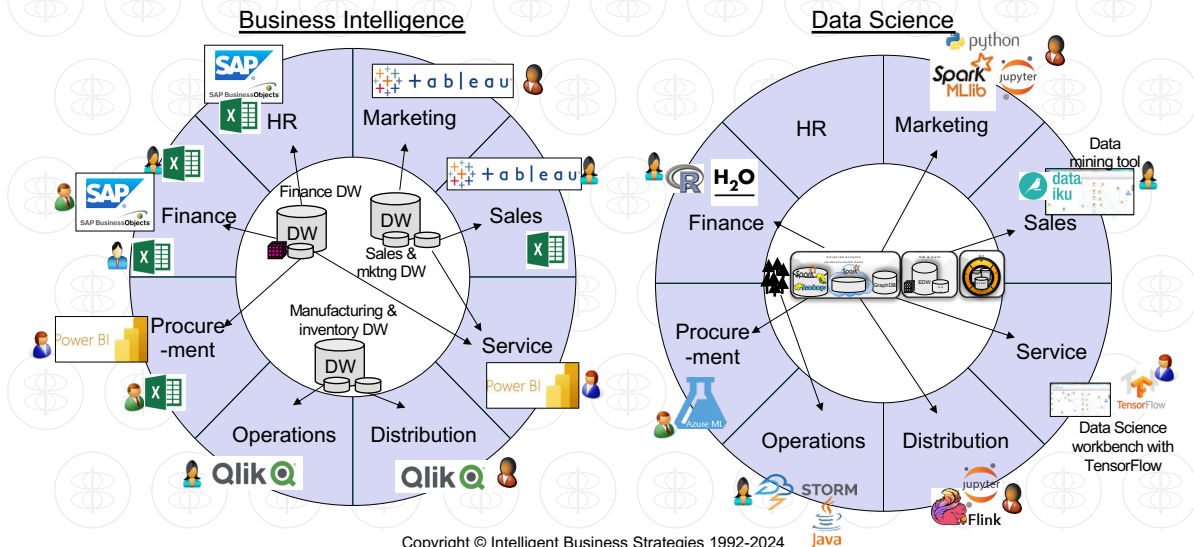






Also, In The Last Decade Many BI Reports, ML Models And AI Bots Have Been Developed Across The Enterprise By Integrating Data Using A Variety Of Tools In Stand-Alone Projects

Most of these stand-alone projects are integrating data in using self-service data preparation tools or code

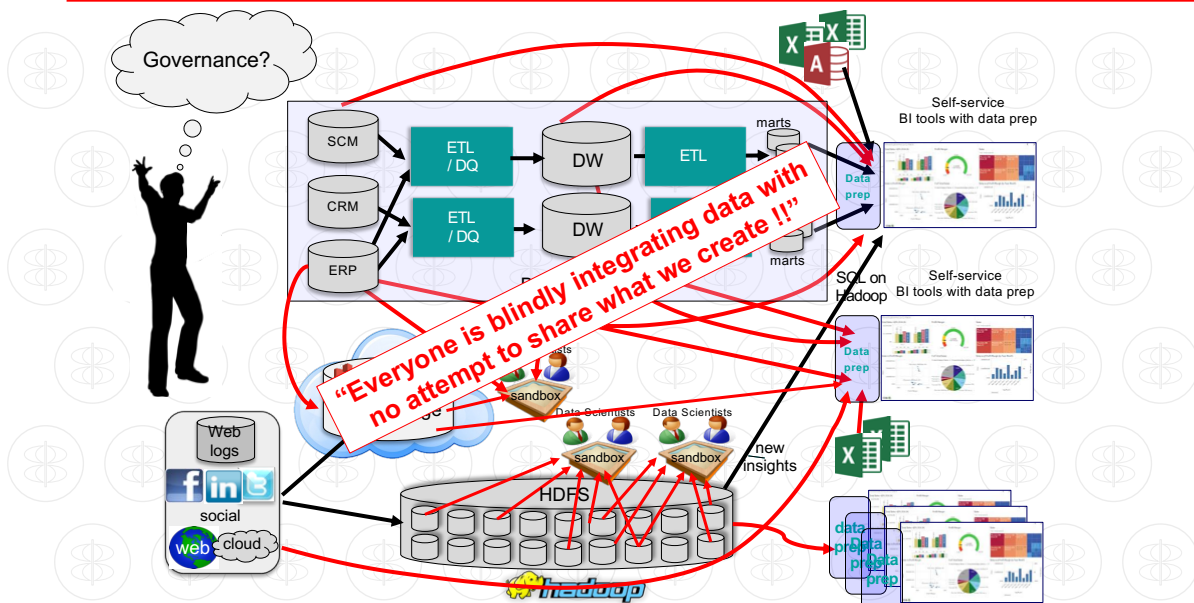


Copyright © Intelligent Business Strategies 1992-2024

11

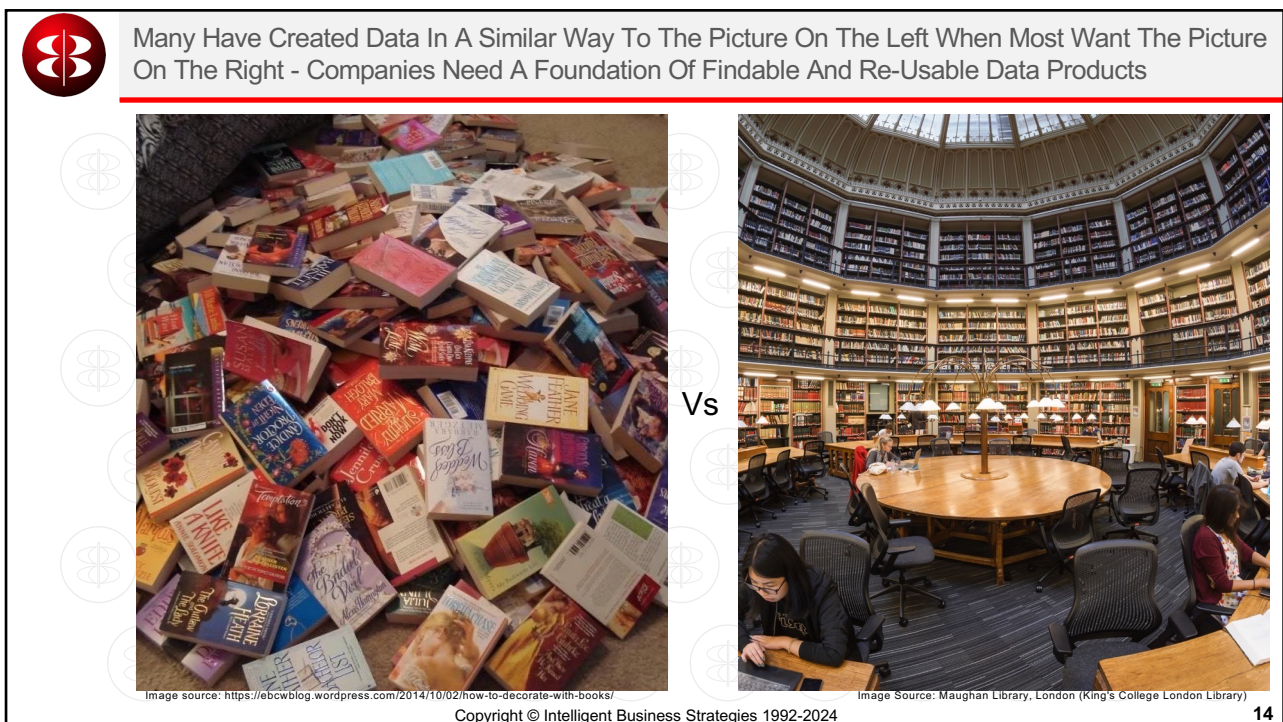
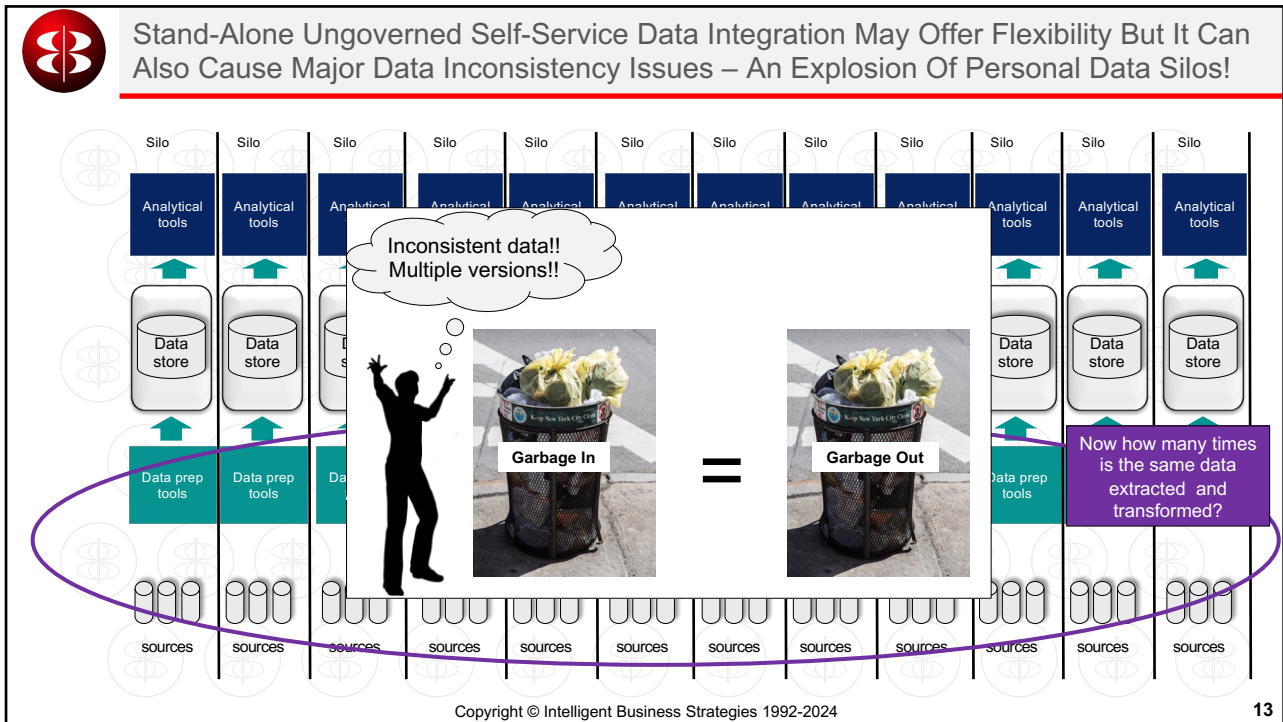


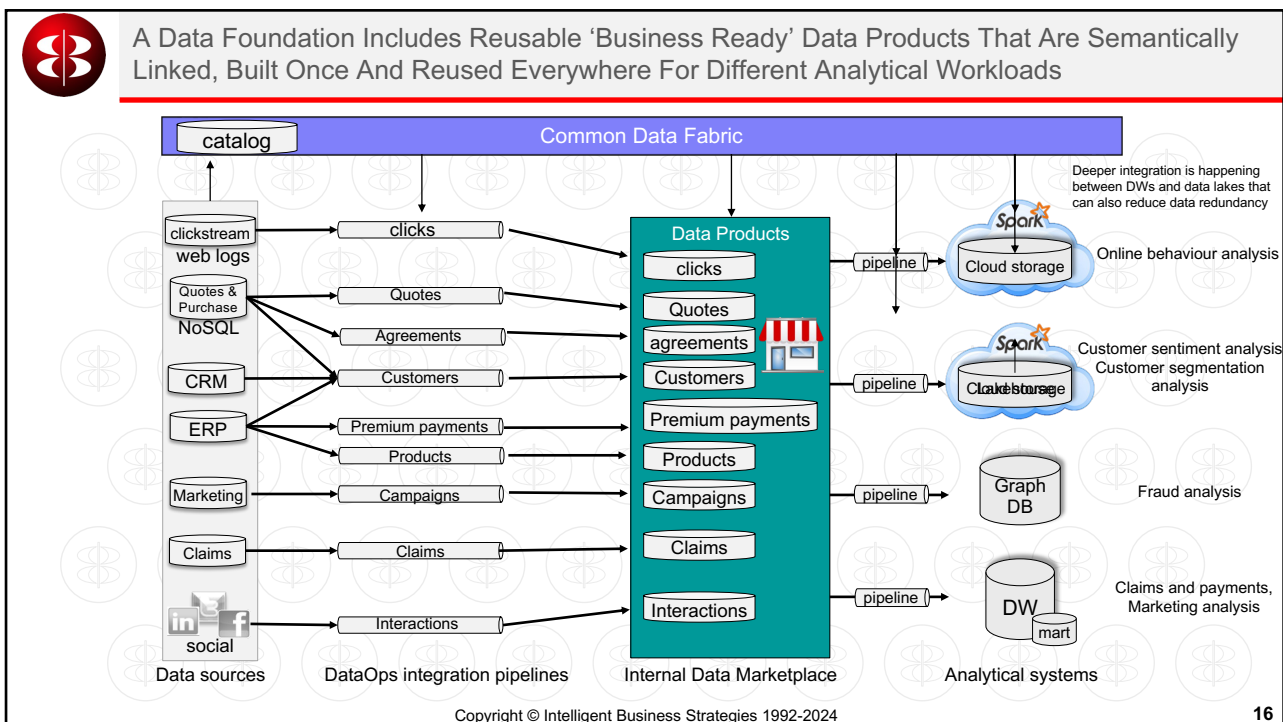
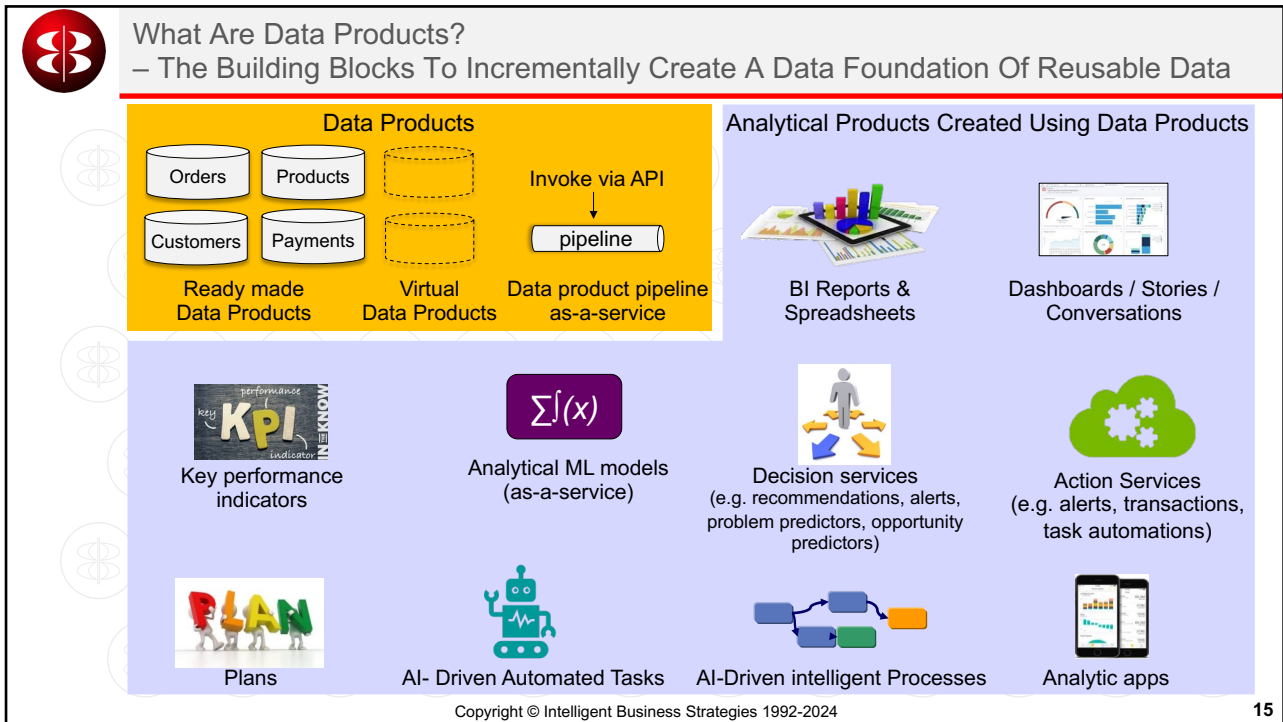
The Problem Is That Ungoverned Self-Service Data Integration Using A Variety Of Tools In Stand-Alone Projects Can Lead To Chaos In The Enterprise



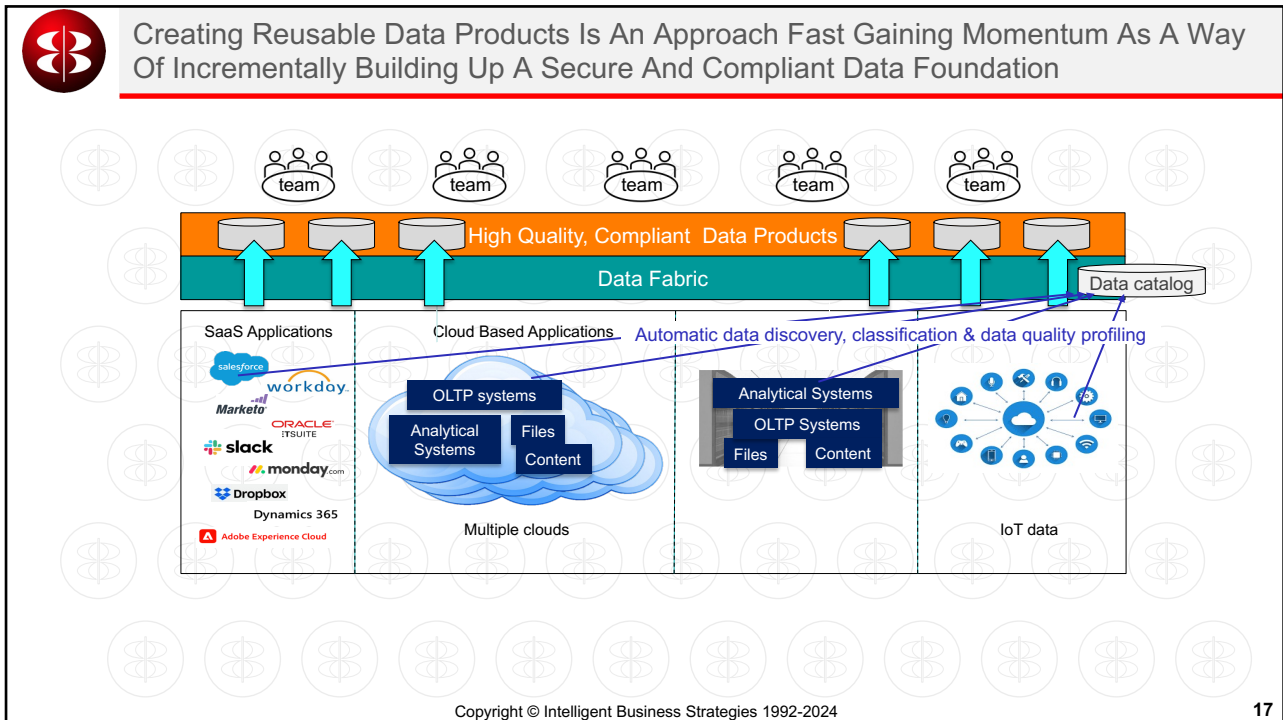
Copyright © Intelligent Business Strategies 1992-2024

12









- 
- Alation
  - Alex Solutions
  - Alteryx Connect
  - Amazon Glue Catalog
  - Apache Atlas (open source)
  - Ataccama ONE Data Catalog
  - Atlan
  - BigID Data Catalog
  - Boomi Atomosphere Data Platform Catalog
  - Cambridge Semantic Anzo Catalog
  - Cloudera Data Platform SDX Catalog
  - Collibra Catalog
  - Databricks Unity Catalog
  - data.world
  - Denodo Catalog
  - Google Cloud Data Catalog
  - Hitachi Vantara Lumada Data Catalog
  - IBM Knowledge Catalog
  - Informatica IDMC Data Governance and Catalog
  - Microsoft Purview
  - Oracle Cloud Infrastructure Data Catalog
  - Orion Governance Catalog
  - Qlik (Talend) Data Catalog
  - Quest (formerly erwin) Data Catalog
  - Rocket Software (formerly ASG) Intelligent Data Catalog
  - SAP Datasphere Catalog
  - SAS Information Catalog
  - Salesforce Tableau Catalog
  - TIBCO Cloud Metadata Catalog
  - Top Quadrant TopBraid EDG Data Catalog
  - Truist (Zaloni) Arena Data Catalog
- Copyright © Intelligent Business Strategies 1992-2024



## Example Data Fabric (Data Management) Platforms Competing To Manage Data Across A Distributed Data Landscape

- Ab Initio
- Alteryx
- AWS Glue and Lake Formation and DataZone
- Ataccama ONE Platform
- Boomi AtomSphere Platform and Data Catalog
- Cambridge Semantics Anzo
- Denodo
- Global IDs Data Ecosystem Evolution Platform (DEEP) Platform
- Google Cloud Dataplex, Data Fusion and Data Catalog
- IBM Cloud Pak for Data (includes IBM Knowledge Catalog), watsonx and StreamSets
- Informatica Intelligent Data Management Cloud (IDMC)
- Infoworks platform
- Microsoft Fabric (includes Azure Data Factory, Synapse Data Engineering, Synapse Data Warehousing, Data Science) and Purview
- Oracle - Enterprise Metadata Management, Data Catalog, Data Quality and Data Integrator
- Qlik (Talend) Data Fabric
- SAS Viya Data Management Platform
- SAP DataSphere (includes SAP DataSphere Catalog)
- Stratio Generative AI Data Fabric
- TIBCO (IBI) Data Management Platform
- Truist Zalon Arena Data Platform

Some vendors include their Data Catalog with their Data Fabric software while others sell their Data Catalog as a separate product

# Data Fabric

Copyright © Intelligent Business Strategies 1992-2024

19



## Trends – Companies Want A Common Data And Analytics Software Platform With Integrated Tools And Shared Metadata To Accelerate Development

Build data products, analytical products and AI-driven applications faster with end-to-end governance

# Single Vendor Data Fabric

Vs

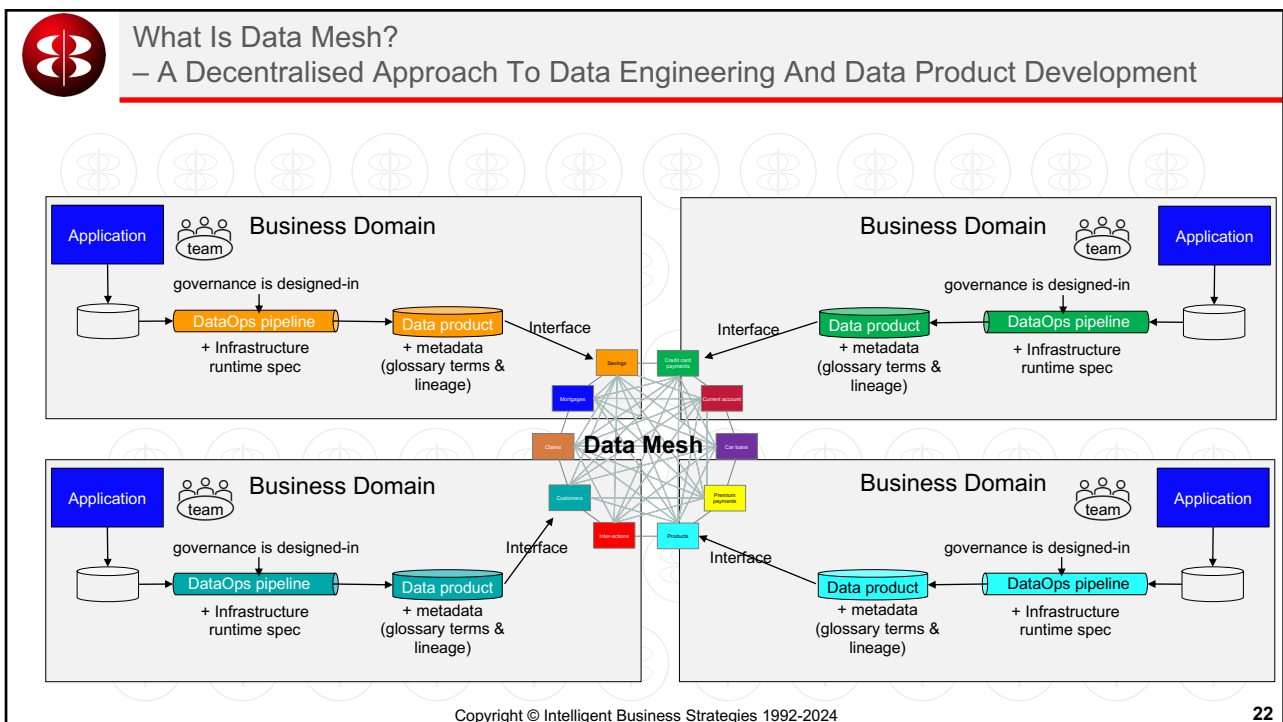
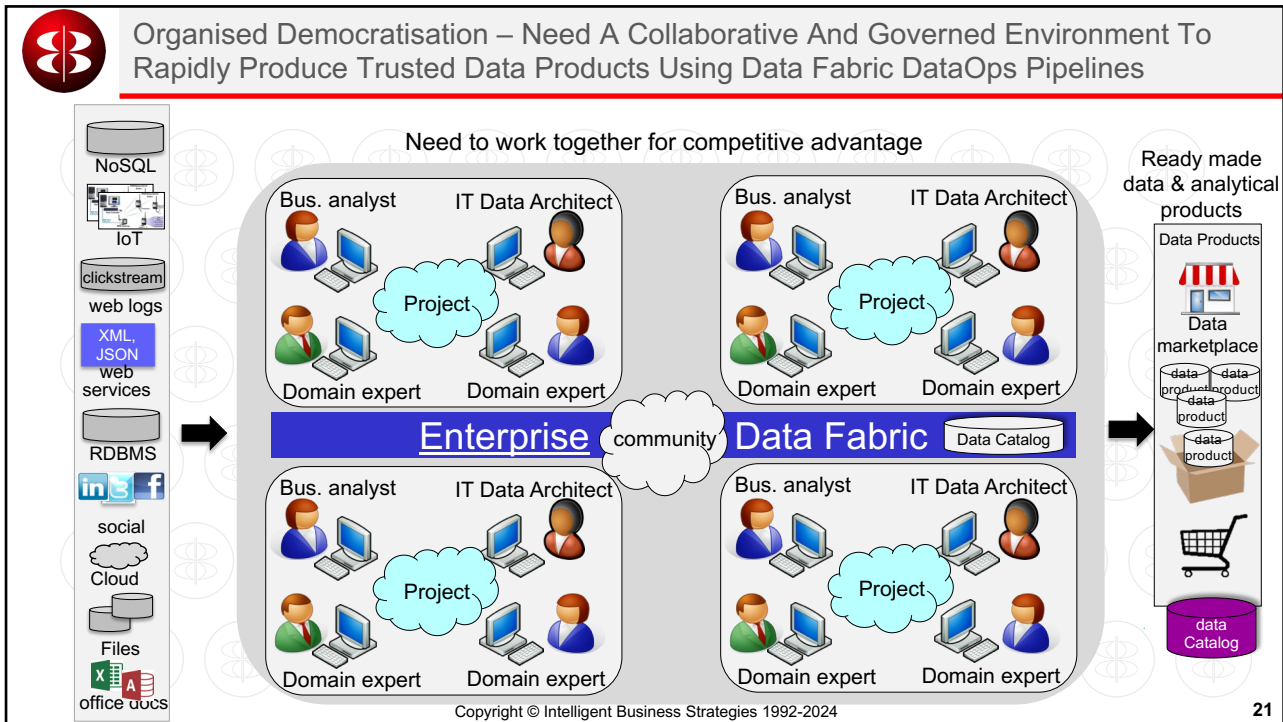
# The Modern Data Stack

- |  |  |
|--|--|
| <ul style="list-style-type: none"> <li>• One platform with integrated services and shared metadata                     <ul style="list-style-type: none"> <li>• Manage, build and deploy data and analytical products</li> <li>• Govern data and AI models across the enterprise</li> <li>• Bundled data catalog or 3<sup>rd</sup> party catalog integration</li> </ul> </li> <li>• Some vendors also bundle integrated BI / ML and AI tooling</li> <li>• Attractive because of no standard to share metadata between tools</li> </ul> | <ul style="list-style-type: none"> <li>• A D&amp;A stack made up of complementary best-of-breed tools from multiple vendors who have partnered</li> <li>• Pre-built integrations across the stack</li> <li>• Formed to compete against single vendor data fabric</li> <li>• Several variations of the modern data stack have appeared on the market, confusing prospective buyers who are unsure on what tools are integrated</li> </ul> |
|--|--|

**Don't make assumptions – avoid overlaps, ensure tools integrate and metadata can be shared**

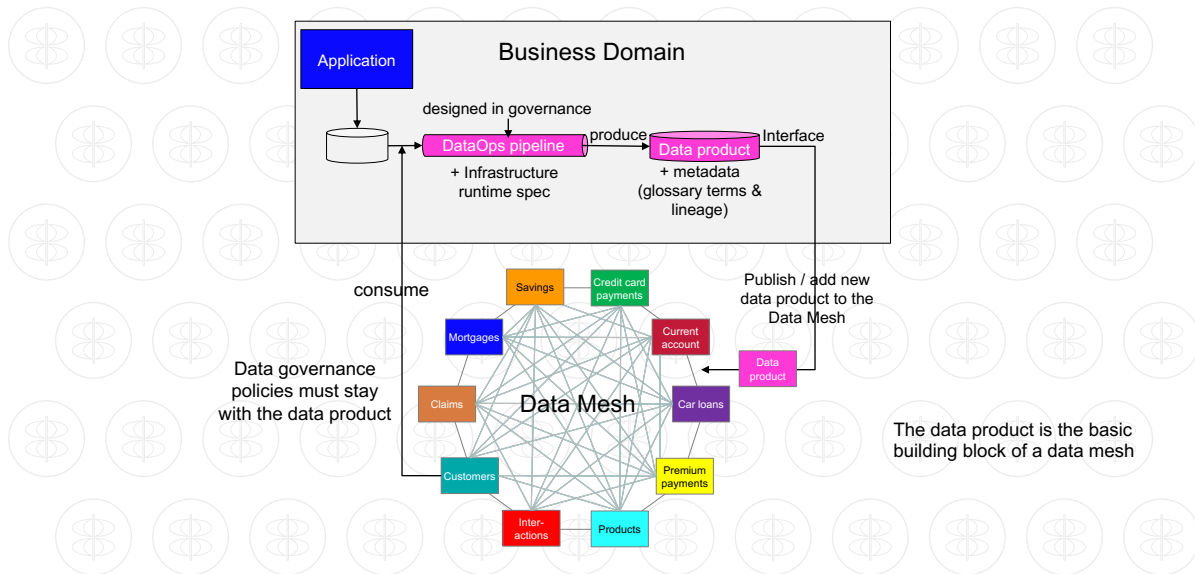
Copyright © Intelligent Business Strategies 1992-2024

20

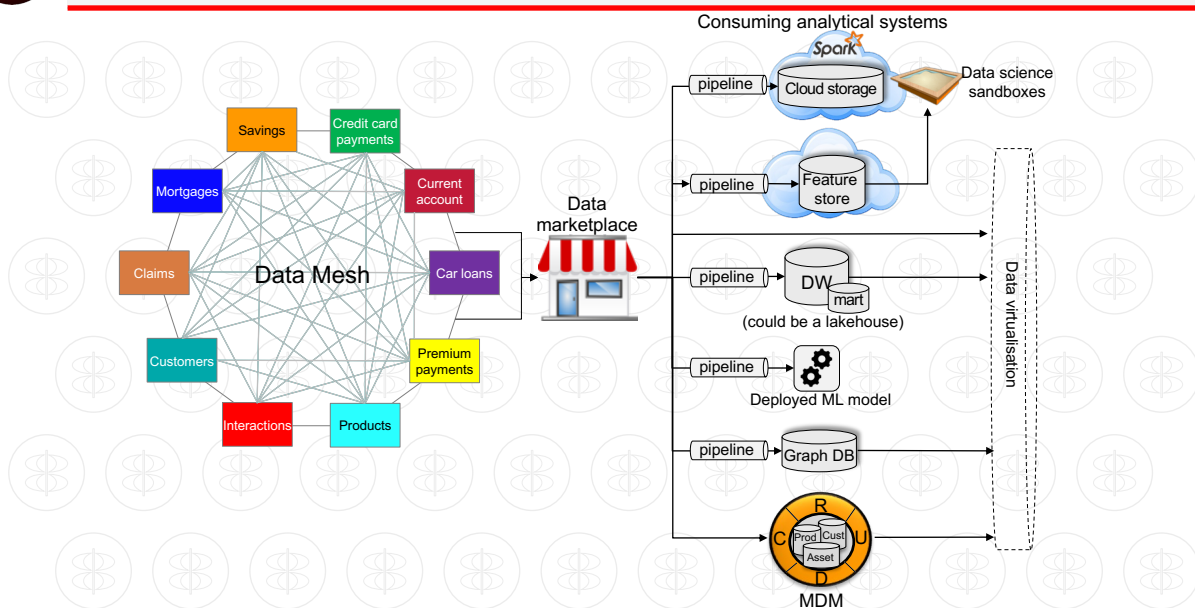




## What Is Data Mesh? – Domains Can Consume Data Products Produced By Other Domains And Then Create New Data Products To Add To A Data Mesh



## Potential Analytical Consumers Of Reusable Data Products In A Data Mesh





## Multiple Architecture Options Have Emerged For Decentralised Creation Of Data Products?

**Multiple approaches to creating data products**

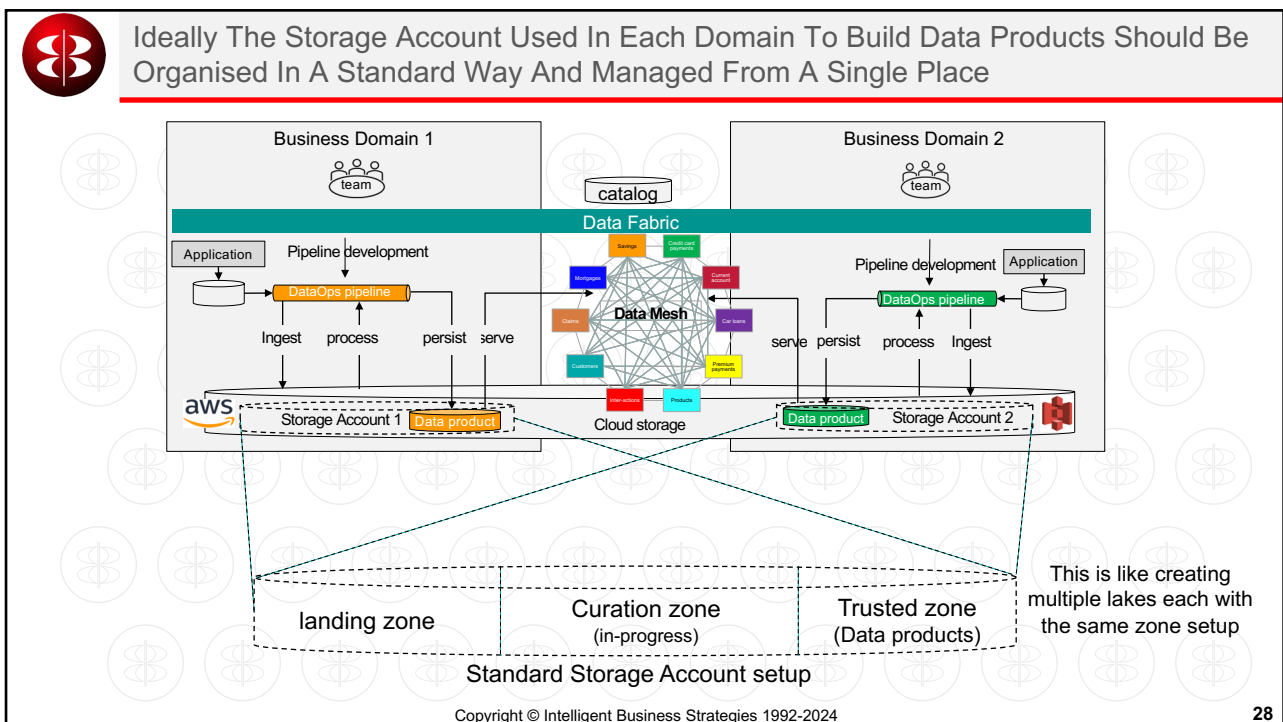
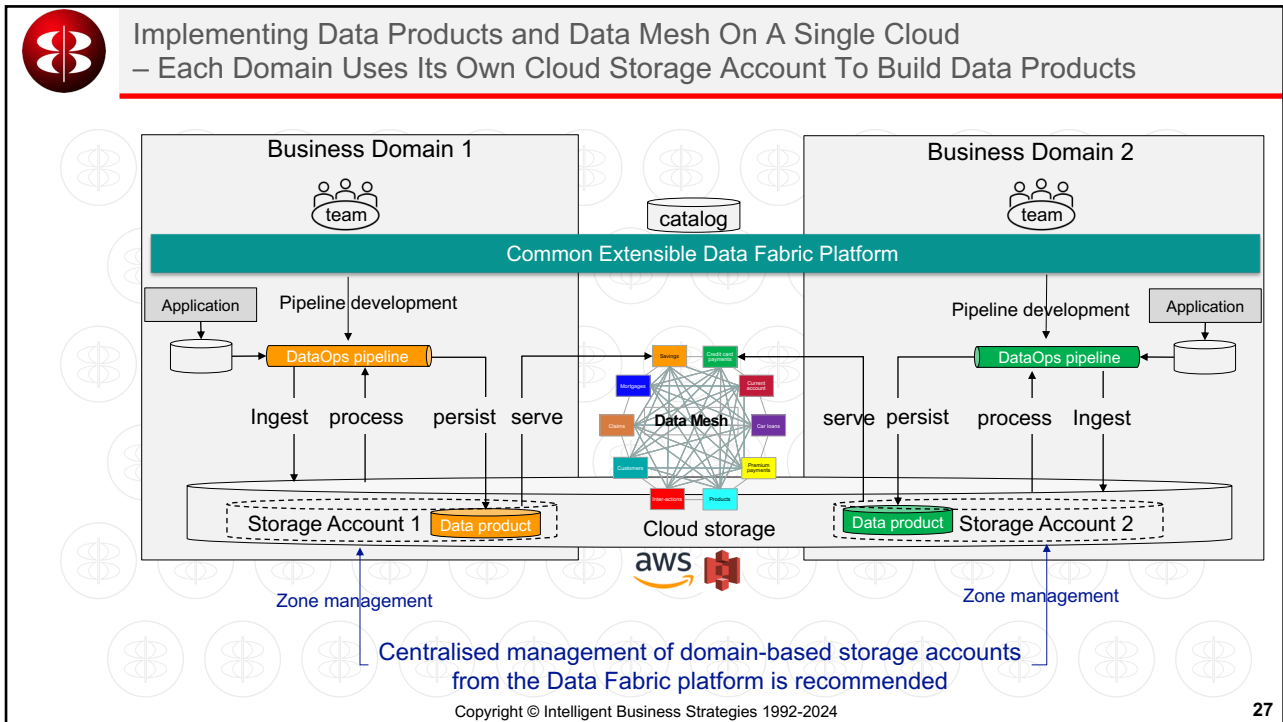
1. One storage account, multiple zones on cloud storage
2. Multiple zoned storage accounts on a single cloud
3. Multiple zoned storage accounts on multiple clouds
4. One or multiple zoned lakehouses (one per domain)
5. Create data products in a data warehouse staging tables
6. Leave data where it is and produce virtual data products using data virtualisation / federated queries
7. Create data products using Kafka topics

Data products need to be semantically linked (by using enterprise wide primary and foreign keys)

Copyright © Intelligent Business Strategies 1992-2024 25

## Creating Data Products In A Central Data Lake To Shorten Time-To-Value

Copyright © Intelligent Business Strategies 1992-2024 26





## Storage Accounts Can Be Managed As Multiple Lakes with Zones In A Decentralised Environment Using Data Fabric Software – E.g. Google Dataplex

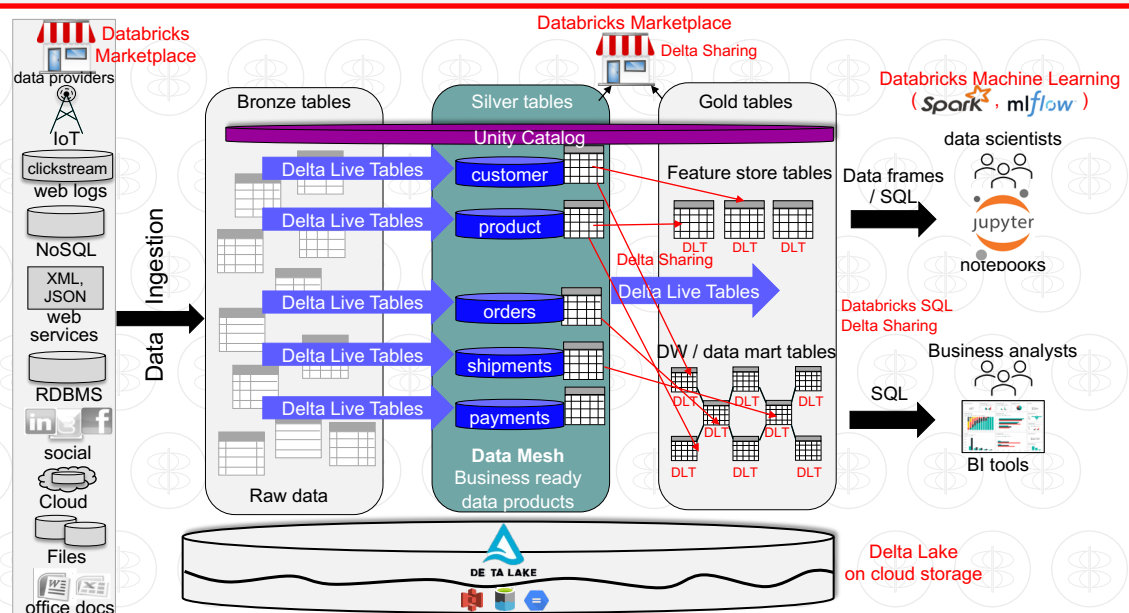
The screenshot shows the Google Cloud Platform interface for a lake named 'Bikeshare'. It displays summary statistics: 4 total zones, 1346 total assets, 0 zones requiring action, and 0 assets requiring action. Below this, the lake details are shown: Lake ID 'bikeshare-lake', Display name 'Bikeshare', Type 'Lake', and Status 'Active'. A table lists the zones:

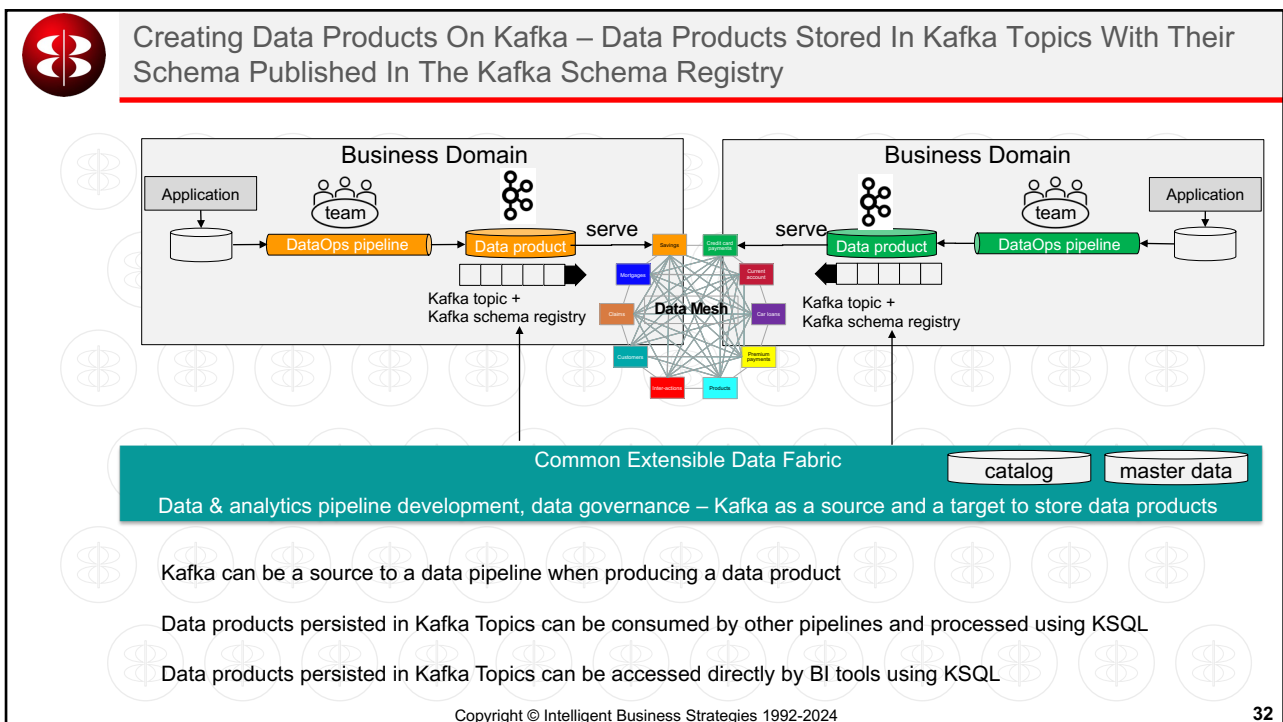
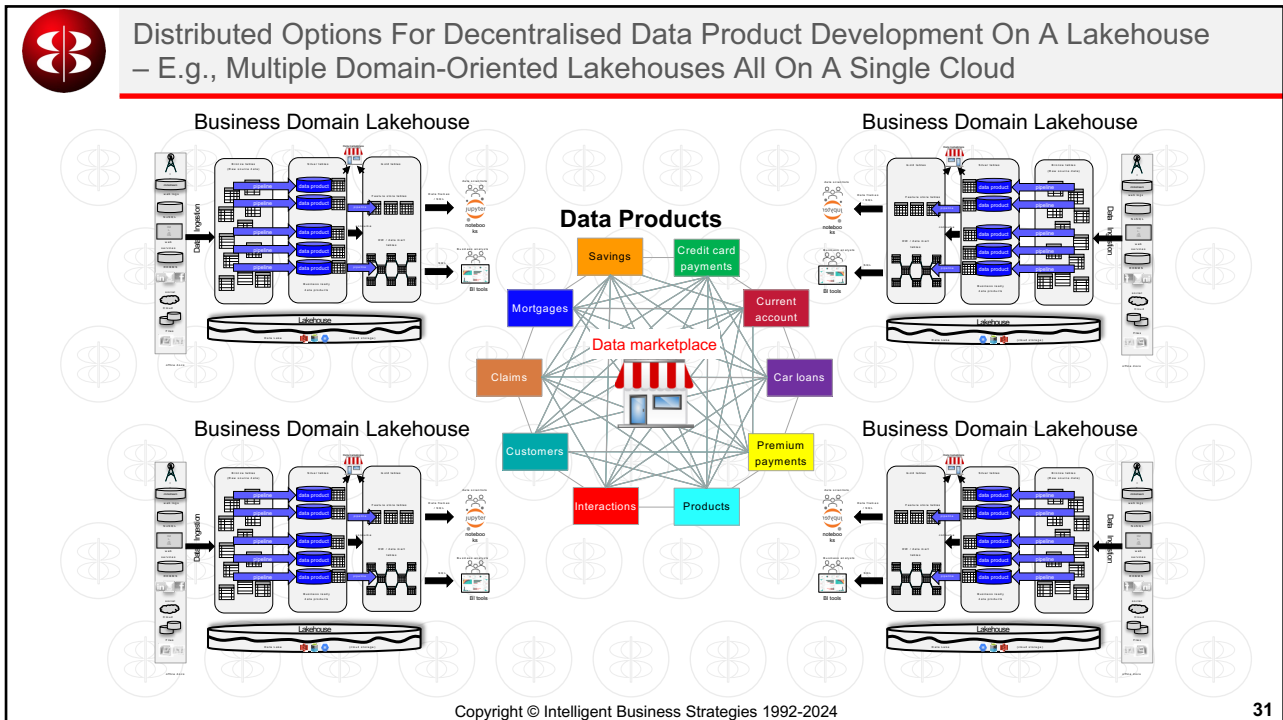
Display name	Type	Status	Assets	Last modified	Label
Landing Zone	Raw	Active	1342	May 26, 2021, 2:34:55 PM	ingestiontemplate: c85ad-secondary...
Raw Zone	Raw	Active	1	May 26, 2021, 10:44:12 AM	
Data Analytics Zone	Curated	Active	2	May 26, 2021, 10:44:12 AM	
Data Science Zone	Curated	Active	1	May 26, 2021, 10:44:12 AM	

You can create a lake per business domain within your organisation and create data zones that map to data readiness and usage (landing zone, raw data zone, data products zone, data science zone, etc.)



## Creating Data Products On A Lakehouse Example – Databricks Lakehouse Platform







### Data Products Persisted In A Kafka Topic - Event Streams Enable You To Record An Immutable History, Consumers Can Replay To Take The Data They Need

**Business Domain**

Application → DataOps pipeline (team) → Data product → Kafka topic + Kafka schema registry (1-9) → Consumer tool / application

KSQL → BI Tool

- Replay enables aggregation
- Schema registry ensures commonly understood definitions
- New version of a data product can be stored in a new topic

Copyright © Intelligent Business Strategies 1992-2024 33

### New Modern Data Architecture Option – Build Data Products In Lakehouse Open Tables On A Multi-Cloud Data Lake With Data Warehouse And Lakehouse Integration

Other engines can directly update and access open tables: presto, Flink, Spark, jupyter

**Extended Data Warehouse DBMS**

External table, Native open table, Native open table, External table

External table mappings, federated query & data pipelines

**Data Fabric**

CSP 2 Lakehouse, CSP 1 Lakehouse, CSP 3 Lakehouse

Cloud storage (data lake), Cloud storage (data lake), Cloud storage (data lake)

Trusted Data Products

**Data Fabric**

SaaS Applications: Salesforce, Marketo, Slack, Monday.com, Dynamics 365, Adobe Experience Cloud

Cloud Based Applications: OLT systems, Analytical Systems, Files, Content, Multiple clouds

On-Premises Systems: Analytical Systems, OLT Systems, Files, Content

Edge Devices: IoT data

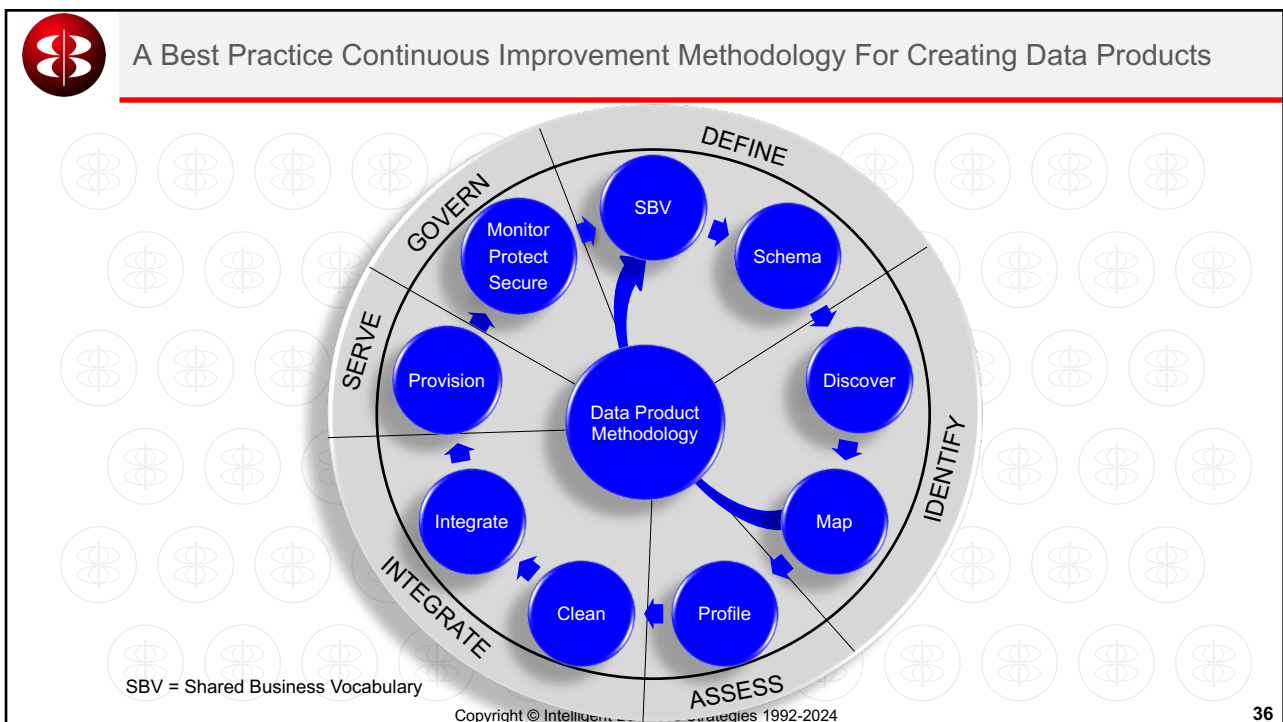
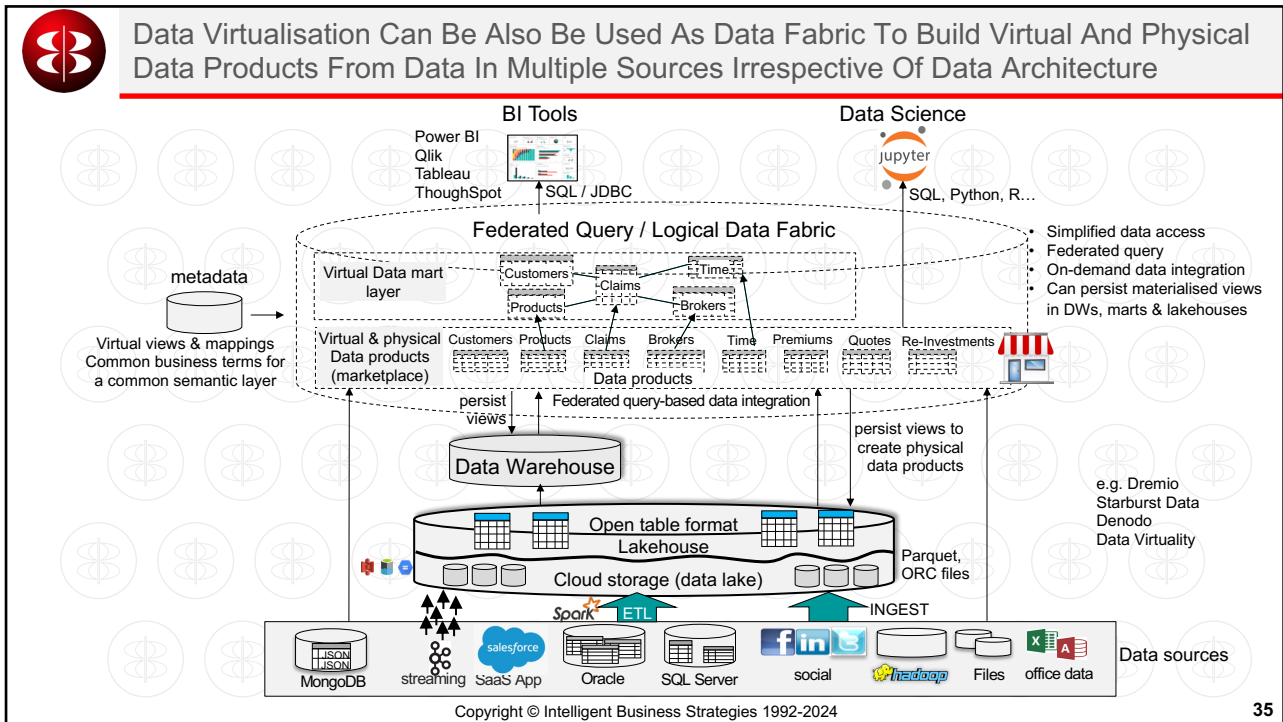
Legend:

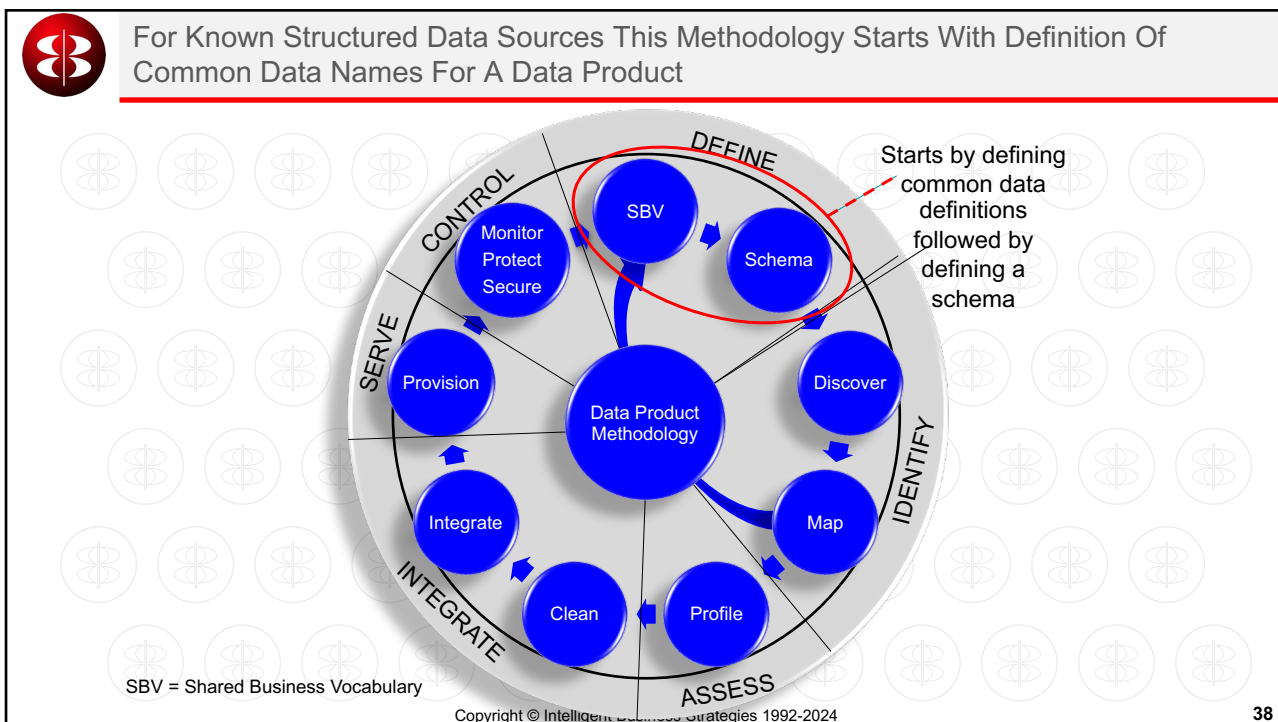
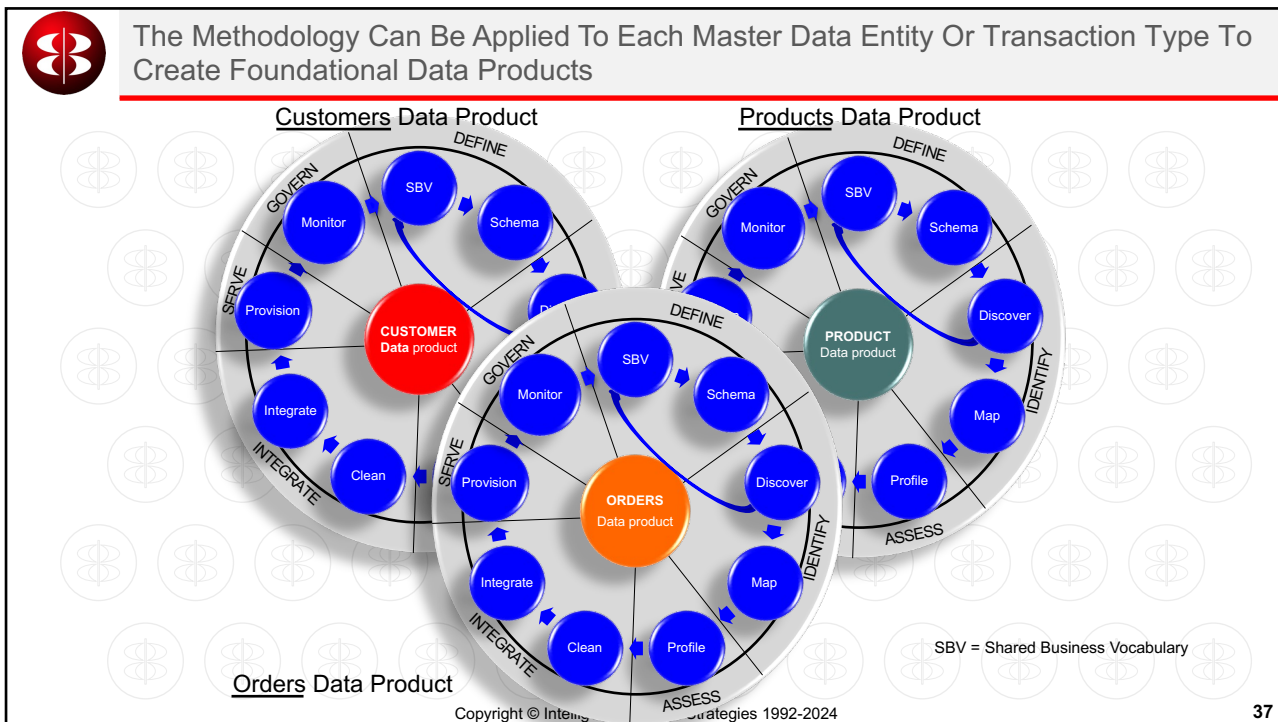
- DBMS proprietary table format
- DBMS external table
- Open table format (Delta Lake, Iceberg, Hudi)

E.g.,

- Google BigQuery, Big Lake & Dataplex,
- Microsoft Fabric & OneLake,
- IBM Db2 & watsonx.data (IBM watsonx.data also bridges to on-premises data)

Copyright © Intelligent Business Strategies 1992-2024 34

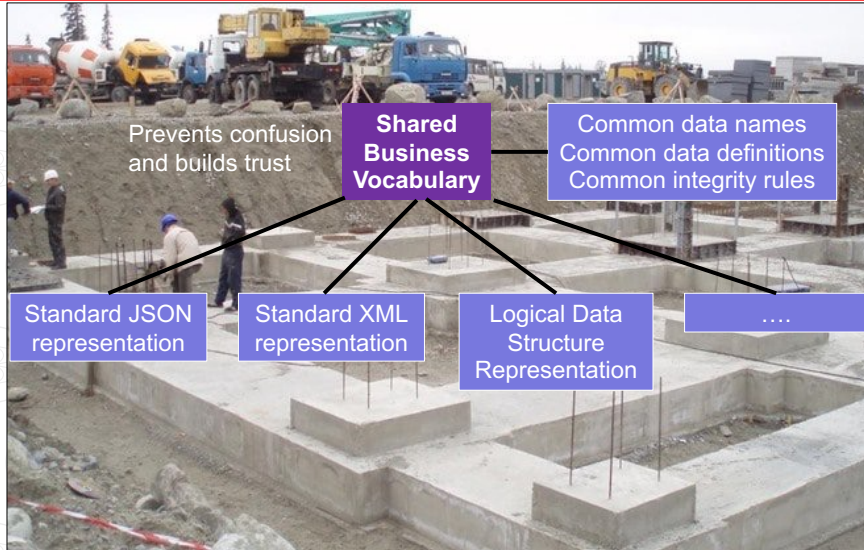






### A Shared Business Vocabulary Is Critical To Creating Data Products To Incrementally Build A Data Foundation, Creating Metrics To Provide Insights And For Data Governance

Data standards and semantics provide consistency



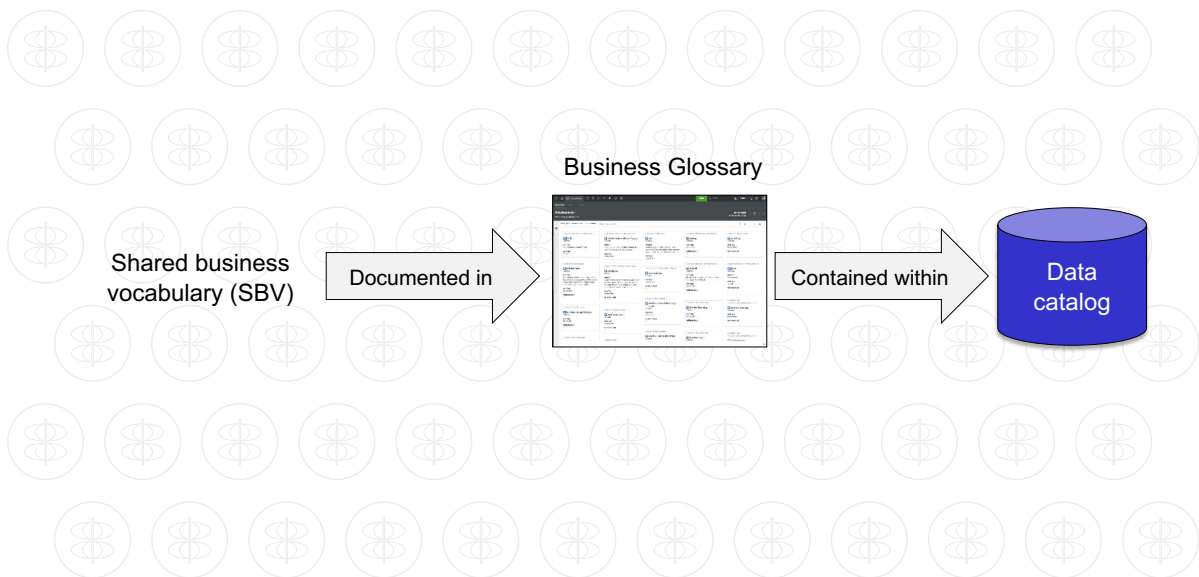
It provides a business lens on your data

Acts as the foundation for sharing data and is fundamental to getting rid of complexity

Copyright © Intelligent Business Strategies 1992-2024



### A Shared Business Vocabulary Is Defined In A Business Glossary Which Is Typically Part Of A Data Catalog



Copyright © Intelligent Business Strategies 1992-2024



### Why Is A Common Business Vocabulary Relevant – We Need To Map Physical Data To Common Business Data Names

Provides a business lens on your data to help

- Find the data needed to create data products
- Govern data more easily

How many data assets will be discovered and catalogued across your data estate?  
How many copies of the same data exist with different data names?

Enterprise Data Fabric Software

Edge devices → gateway → Sensor Data → On-Premises → Data → Azure → aws → Google Cloud → Data catalog

Automatic data discovery & classification

Business Glossary

Copyright © Intelligent Business Strategies 1992-2024 41

### Data Products Created In A Data Mesh Should Be Defined Using Business Data Names Documented In A Business Glossary

Application → DataOps pipeline (governance must be designed-in, + Infrastructure runtime spec) → Data product + metadata (glossary terms & lineage) → Interface → Data Mesh

Business Glossary

A business glossary is typically inside a data catalog

Copyright © Intelligent Business Strategies 1992-2024 42



## Business Glossary Products Provide Users With Access To A Common Business Vocabulary And Maintenance Of The Vocabulary

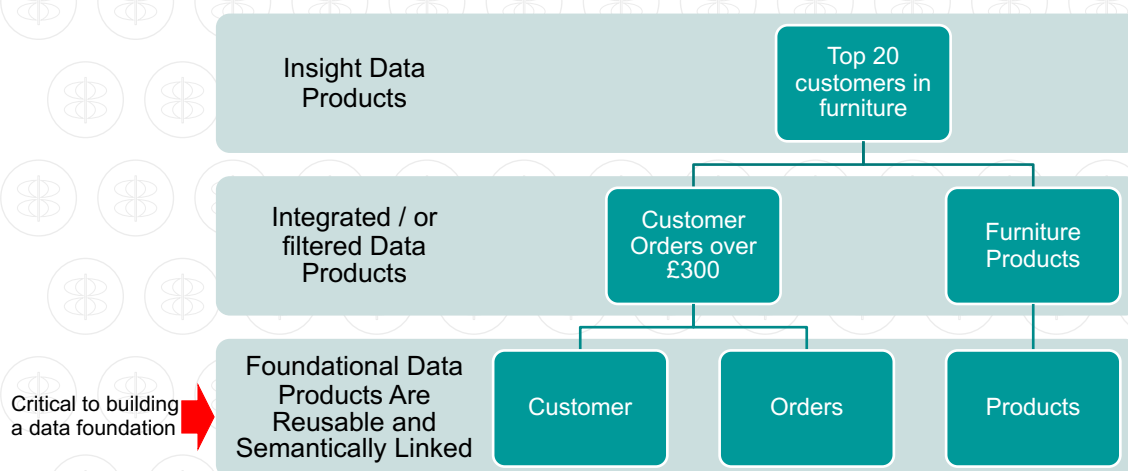
### Business Glossary Product Examples:

- Alteryx Connect Glossary
- Rocket Software (formerly ASG) Business Glossary
- Collibra Business Glossary
- IBM Watson Knowledge Catalog Business Glossary
- Informatica Axon Business Glossary (integrates with Enterprise Data Catalog)
- Microsoft Purview Business Glossary
- Oracle Cloud Infrastructure Data Catalog
- SAP DataSphere Catalog and semantic layer
- SAS Business Data Network
- Qlik (Talend) Data Catalog Business Glossary
- TIBCO Cloud Metadata
- Top Quadrant TopBraid EDG Business Glossary



## Foundation Data Products Are Critical In Any Business – There Are Different Levels Of Data Products

All data products must have an ID



**Basic Mechanisms Can Help You Get Started In Creating Foundational Data Products – What Are The Core Master Data Entities And Transaction Types In Your Organisation?**

Insurance		Banking	
Master data entities	Transaction types	Master data entities	Transaction types
Customer	Quote	Customer	Open / Close Account
Agreement	Policy issuance	Product	Deposit
Product	Premium payment	Account	Withdrawal
Broker	Claim	Employee	Money Transfer
Beneficiary	Claim payment	Counterparty	Loan payment
Supplier	Renewal	Legal entity	Credit card transaction
Employee	End of term / settlement	Supplier	FX trade
Risk (Car, Property...)			Equity trade
			Derivative trade

Each of these master data entities and transaction types can become a foundational data product

45

**Popular Master Data Entity Examples**

- Customer
- Product
- Employee
- Supplier
- Asset
- Location
- Financial Chart of Accounts

These are master data products

The problem is they are cross domain

They are needed in a Data Mesh

1. Master data needs to be owned
2. Master data pipelines need to be developed
3. Master data products created, stored in an MDM system and published in a data marketplace

The same data fabric technology can be used to create master data products and all other data products

Copyright © Intelligent Business Strategies 1992-2024

46



## Some Vendors Provide Pre-Built Vocabularies, Key Performance Indicators, Reference Data And Classifications For Different Vertical Industries – IBM Knowledge Accelerators

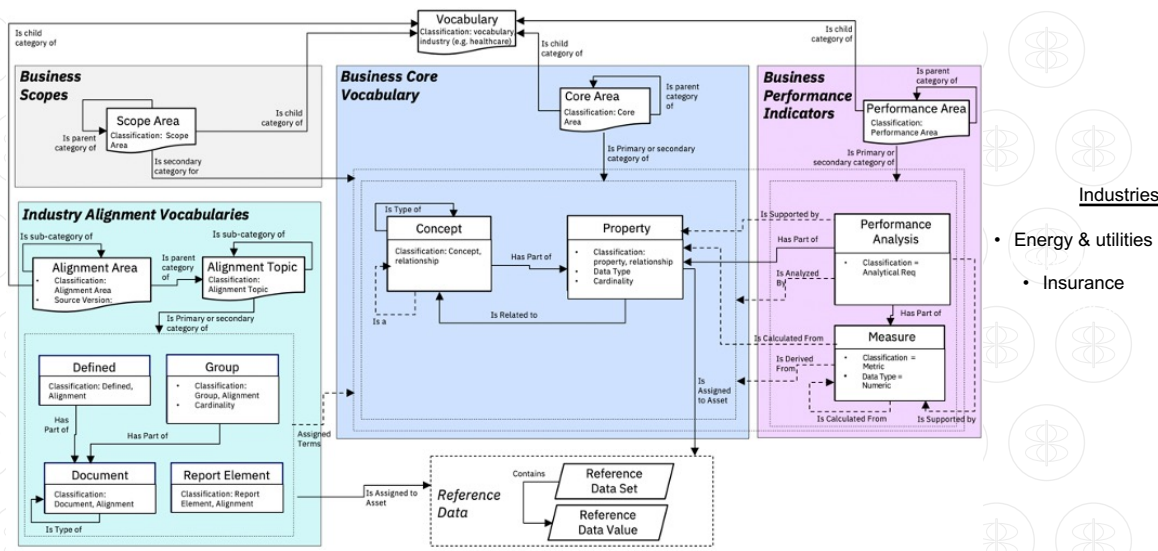


Image source: <https://community.ibm.com/community/user/dataops/blogs/karen-madera1/2020/08/07/building-an-extensible-business-vocabulary>

Copyright © Intelligent Business Strategies 1992-2024



## Some Vendors Provide Pre-Built Conceptual Data Models And All Their Attributes To Get You Started Quickly, e.g., Collibra Integrates With Schema.Org, Microsoft CDM

Source: <https://learn.microsoft.com/en-us/common-data-model/>

Copyright © Intelligent Business Strategies 1992-2024



## SAP One Domain Model (ODM) – Business Objects Are Automatically Mapped From SAP And Non-SAP Solutions To The Appropriate One Domain Model Fields

**Consistent master data across all SAP apps**

**Shared context through ODM**

As the harmonized domain model for objects that are distributed throughout the different applications, SAP One Domain Model provides a basis for a consistent view on master data across the entire hybrid landscape.

By mapping objects to a central domain model, SAP One Domain Model enables applications to speak different languages, signs configuration and transactional data, and sets the generic.

Source: SAP

es 1992-2024

49

## A Good Way To Identify Potential Data Products Is To Create A Business Data Concept Model

- Data concept modelling is a very important skill if decentralised development is to succeed
- Identify the data concepts, properties and relationships and construct a data concept model
- It is good practice to highlight all master and transaction data concepts in your data concept model

### Business Data Concept Model Example

Some vendors provide complete pre-built definitions for business data entities and all their attributes to get you started quickly

Copyright © Intelligent Business Strategies 1992-2024

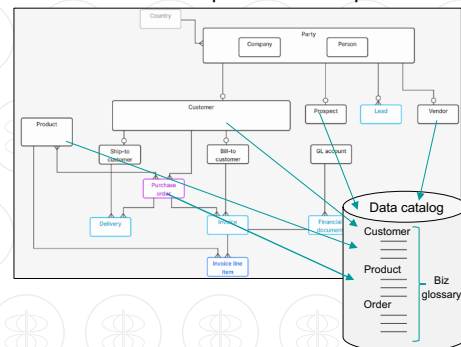
50



## Steps To Creating Data Products – From Data Concept Model To Data Marketplace

1. Identify the data concepts, properties and relationships and construct a data concept model
2. The data concepts become the 'skeleton data entities' in the common vocabulary
3. Each data concept and its attributes should be defined in the business glossary as a data entity with a data owner
4. Use the catalog to discover the data for each data entity in underlying data stores across the data landscape
5. Design DataOps component-based pipelines to create the data products with common vocabulary data names
6. Publish all data products in a data marketplace

Data Concept Model Example



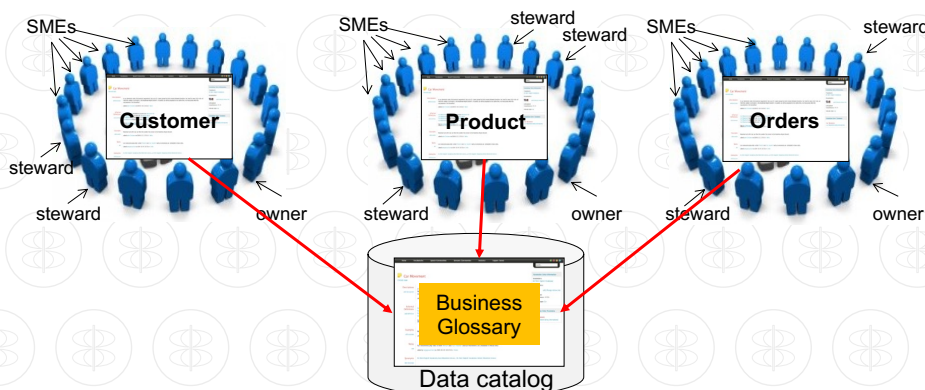
Copyright © Intelligent Business Strategies 1992-2024

51



## Multiple Domain And Cross Domain Oriented Communities Are Likely To Be Involved In Defining Common Data Names And Definitions For Data In A Business Glossary

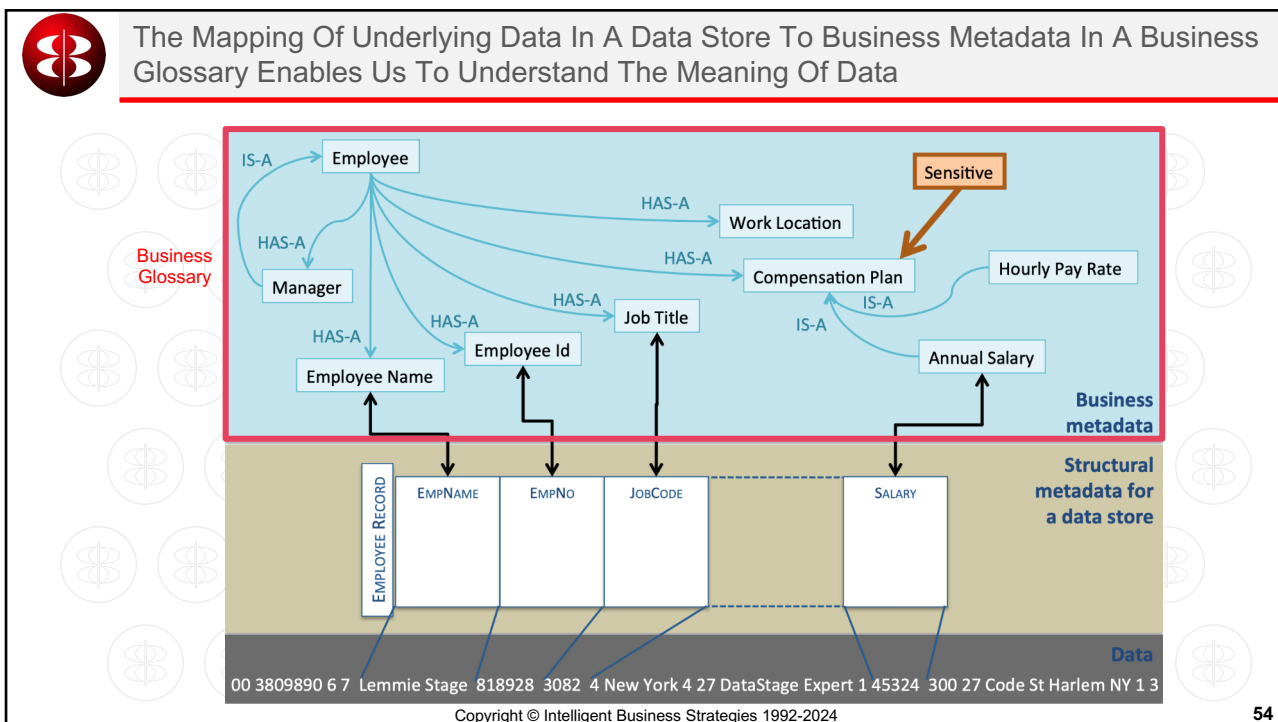
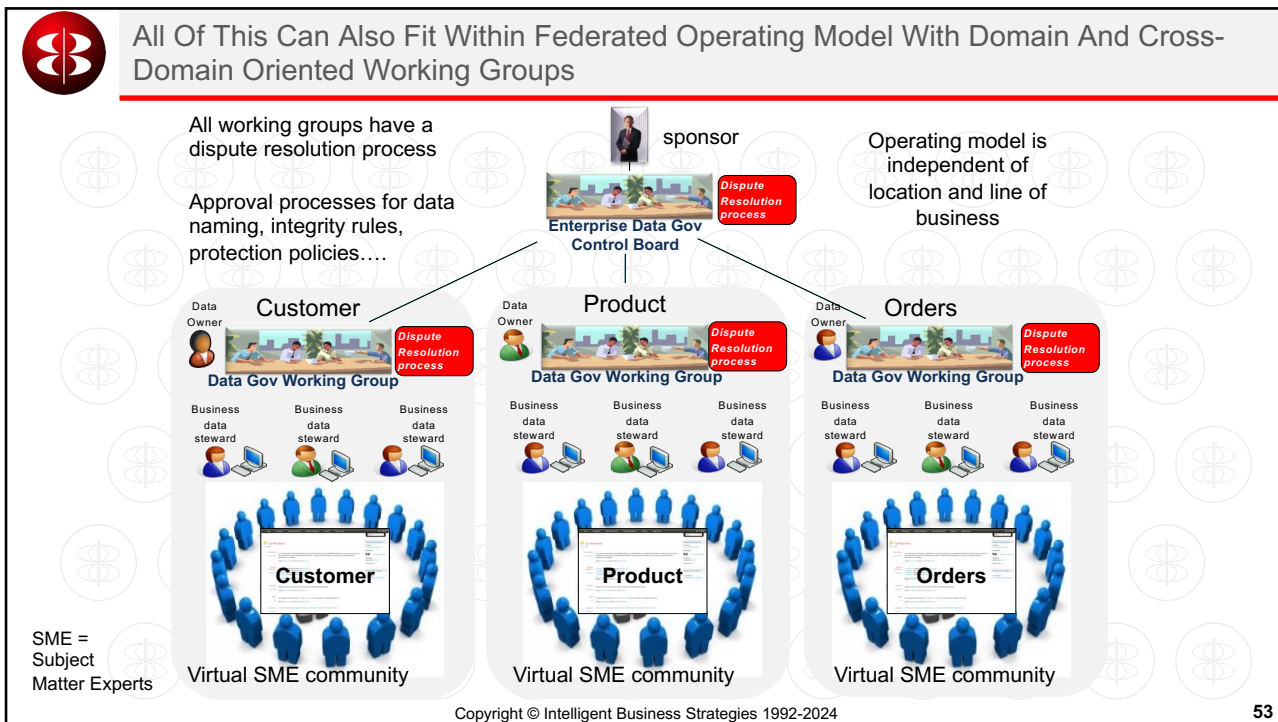
Communities may include a mix of domain-oriented subject matter expert (SME) business users and IT

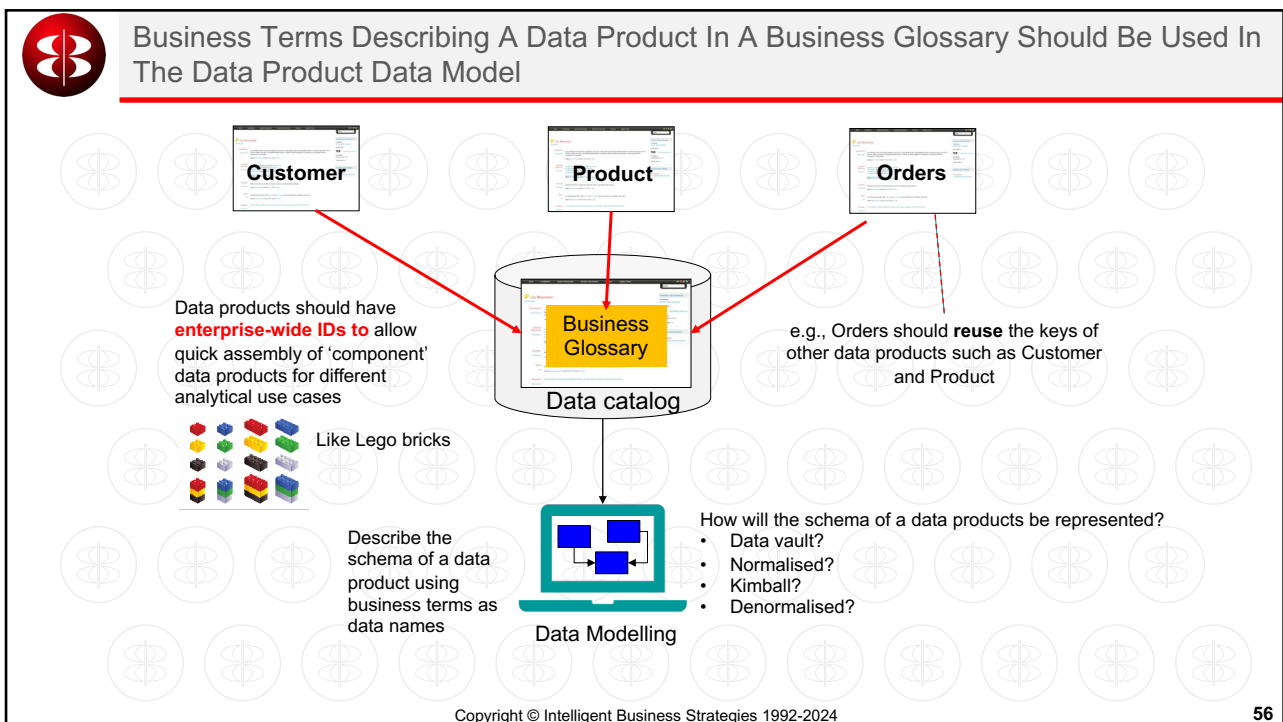
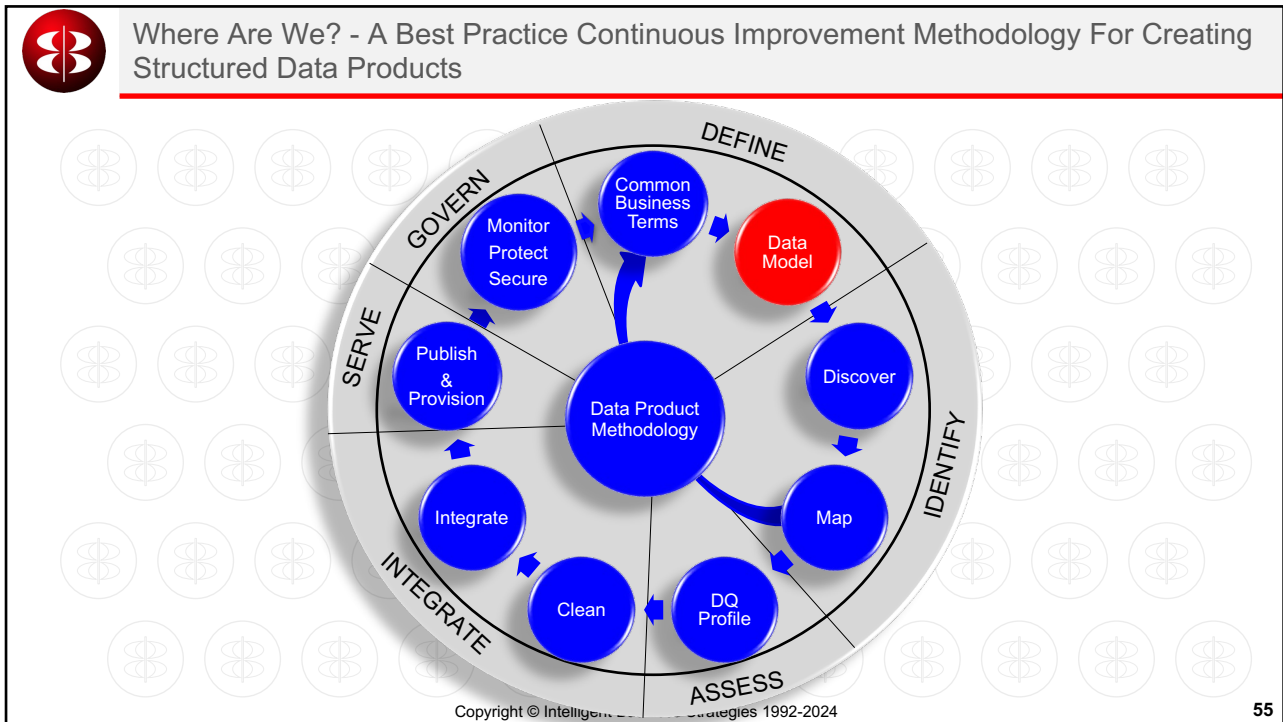


**Master data products are cross-domain** and so a cross-domain oriented community is needed to define the vocabulary of these data

Copyright © Intelligent Business Strategies 1992-2024

52





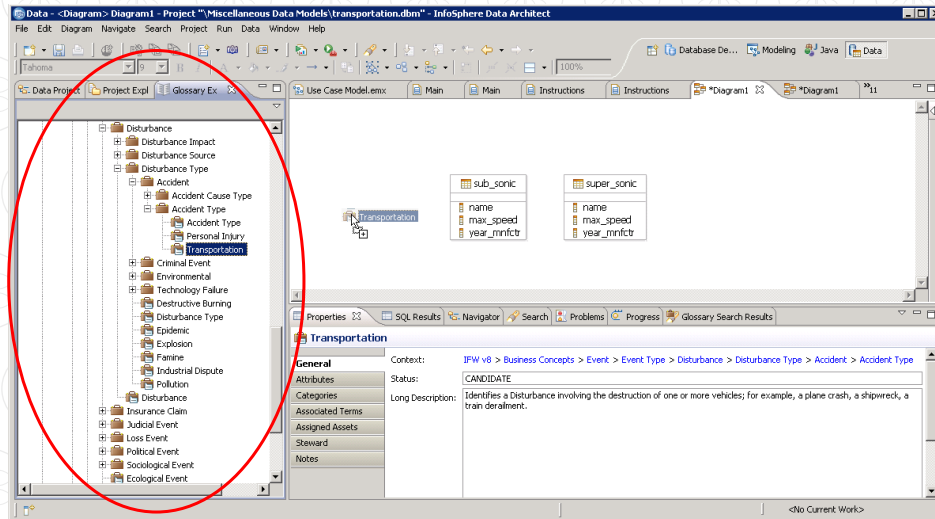




## Construct A Schema Of Your Data Product Using Shared Business Vocabulary Business Data Names In Your Data Product Schema

Business Glossary Eclipse plug-in allows data modellers to use and re-use common business definitions as they model data

Product Example:  
IBM Data Architect



Copyright © Intelligent Business Strategies 1992-2024

57



## Data Product Schema Are Equivalent To Semantically Linked Data Model Components And Become Building Blocks That Can Be Assembled Into Specific Kinds Of Data Model

### Normalised Data Model in Third Normal Form

Data model component

Students			
ID	First Name	Last Name	Marital Status
1	Kevin	Drumm	Single
2	Murvin	Drake	Single
3	John	Jones	Single
4	Sally-Jane	Jones	Single
5	David	Smith	Married

Data model component

StudentCourse		
ID	Course Title	Grade
1	Computer Science	A
1	Mathematics	B
1	Physics	C
2	Physics	B
2	Chemistry	C
3	Music	C
4	Biology	A
4	Economics	B
5	Mathematics	C
5	Physics	D

Courses			
Course Title	Fee	Qualification	Teacher
Computer Science	£2000	Advanced Level	1
Mathematics	£2500	Advanced Level	2
Physics	£1800	Advanced Level	3
Chemistry	£1800	Advanced Level	4
Music	£1200	Diploma	5
Biology	£1000	Certificate	6
Economics	£1500	Diploma	7

Teachers	
Teacher ID	Teacher Name
1	Miss Lovelace
2	Mr Pascal
3	Mr Einstein
4	Mr Bunsen
5	Miss Holiday
6	Mr Darwin
7	Mr Keynes

Image Source: [https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.youtube.com%2Fwatch%3Fv%3D\\_K7fcQowY8&psig=AOvVaw2dh-v2YRKQVnR5g4tpoR&ust=1693406044349000&source=images&cd=vfe&opi=89978449&ved=0CAAOQjRqFwoTCKDLzqWLgoEDFQAAAAAAdAAAAABAD](https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.youtube.com%2Fwatch%3Fv%3D_K7fcQowY8&psig=AOvVaw2dh-v2YRKQVnR5g4tpoR&ust=1693406044349000&source=images&cd=vfe&opi=89978449&ved=0CAAOQjRqFwoTCKDLzqWLgoEDFQAAAAAAdAAAAABAD)

Copyright © Intelligent Business Strategies 1992-2024

58



## Using 3NF<sup>1</sup> Or BCNF<sup>2</sup> As A Data Modelling Technique For Data Products

- Advantages
  - Data is normalised and so no data redundancy
  - Self-contained 'neutral' design
    - All attributes are functionally dependent on the primary key with no transitive functional dependencies
  - Good fit for master data products to be stored in an MDM system as this data can be updated
- Disadvantages
  - Will likely need to be transformed for analytical use
    - Deconstructed into hubs, links and satellites if consuming it to create a data vault based DW
    - Denormalised if consuming it to create a Kimball DW or a data mart

Normalised Data Model in Third Normal Form

Data model component				Data model component						
Students				Courses						
SID	First Name	Last Name	Marital Status	Course Title	Fee	Qualification	Teacher			
1	Kevin	Drum	Single	Computer Science	A	Computer Science	E200	Advanced Level	1	Miss Lovelace
2	Murvin	Drake	Single	Mathematics	B	Mathematics	E250	Advanced Level	2	Mr Pascal
3	John	James	Single	Physics	C	Physics	E180	Advanced Level	3	Mr Einstein
4	Sally Jane	James	Single	Physics	B	Chemistry	E180	Advanced Level	4	Mr Bunton
5	David	Smith	Married	Chemistry	C	Music	E200	Diploma	5	Miss Holiday
				Music	C	Biology	E100	Certificate	6	Mr Darwin
				Biology	A	Economics	E100	Diploma	7	Mr Keynes
				Economics	B					
				Mathematics	C					
				Physics	D					

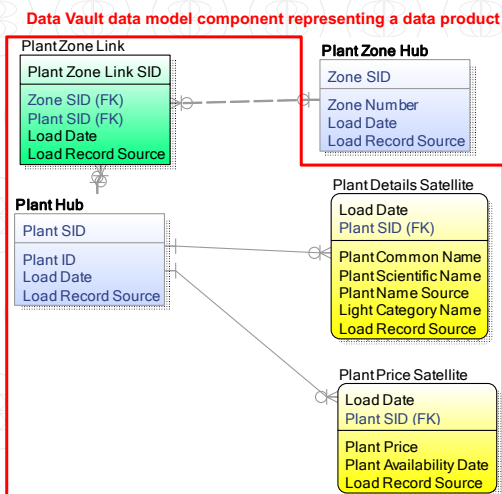
Image Source: [https://www.google.com/url?sa=i&url=https://3A%2F%2Fwww.youtube.com%2Fwatch%3Fv%3D\\_K76FQqay8&psis=AQVWaw2h-vzYR0QVhmr5Sg4t6or&ust=169340614434900&source=images&cd=rv16&pr=89978448&ved=0CAUQKqFwo1CKD.LzVLgE2FC&AA&AAAAABQD](https://www.google.com/url?sa=i&url=https://3A%2F%2Fwww.youtube.com%2Fwatch%3Fv%3D_K76FQqay8&psis=AQVWaw2h-vzYR0QVhmr5Sg4t6or&ust=169340614434900&source=images&cd=rv16&pr=89978448&ved=0CAUQKqFwo1CKD.LzVLgE2FC&AA&AAAAABQD)

<sup>1</sup> 3NF = Third Normal Form  
<sup>2</sup> BCNF – Boyce / Codd Normal Form (a stronger version of third normal form than Codd's original definition)



## Data Products Are Equivalent To Semantically Linked Data Model Components And Become Building Blocks That Can Be Assembled Into Specific Kinds Of Data Model

Linstedt Data Vault



Easy to change  
 Hubs, links and satellites can be hidden by a view

Kimball Star Schema

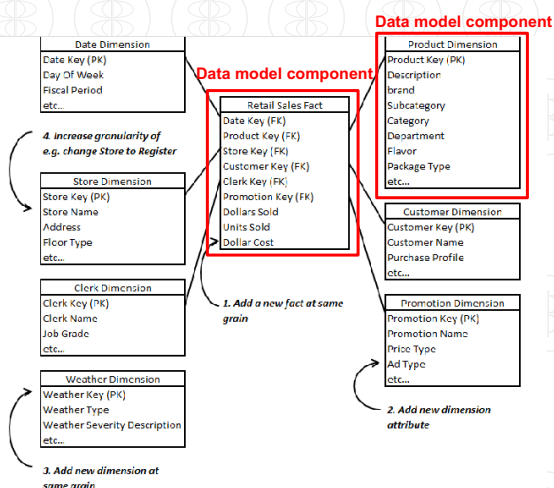


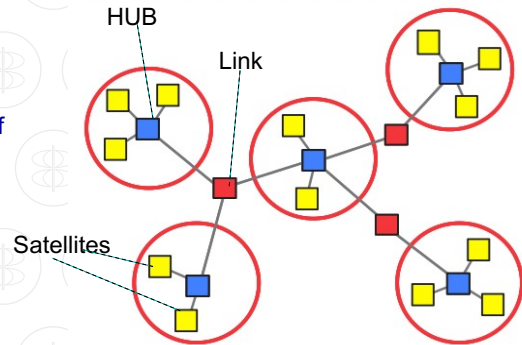
Image Source: A Dimensional Modelling Manifesto, Ralph Kimball, August 1997



## Using Data Vault As A Data Modelling Technique For Data Products – Hubs, Links And Satellites Represent Data Entities That Can Be Turned Into Data Products

- Advantages
  - Hubs, links and satellites are already normalised
  - Data warehouse ready data products
  - Very easy to change with minimal impact
    - Adding a column can be a new satellite
  - Can create views to provide an integrated view of the data product
  - Access the data via a business view
- Disadvantages
  - Data products built for analytical use only
  - Harder to learn for citizen data engineers
- Other
  - Needs denormalised if consuming into a data mart

Linstedt Data Vault



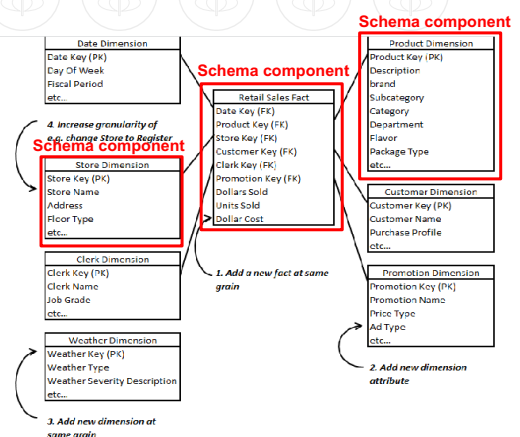
- HUB – Load ELT processes – Insert only
- LINK – Load ELT processes – Insert only
- SAT(elite) – Load ELT process – Insert (and Update) only

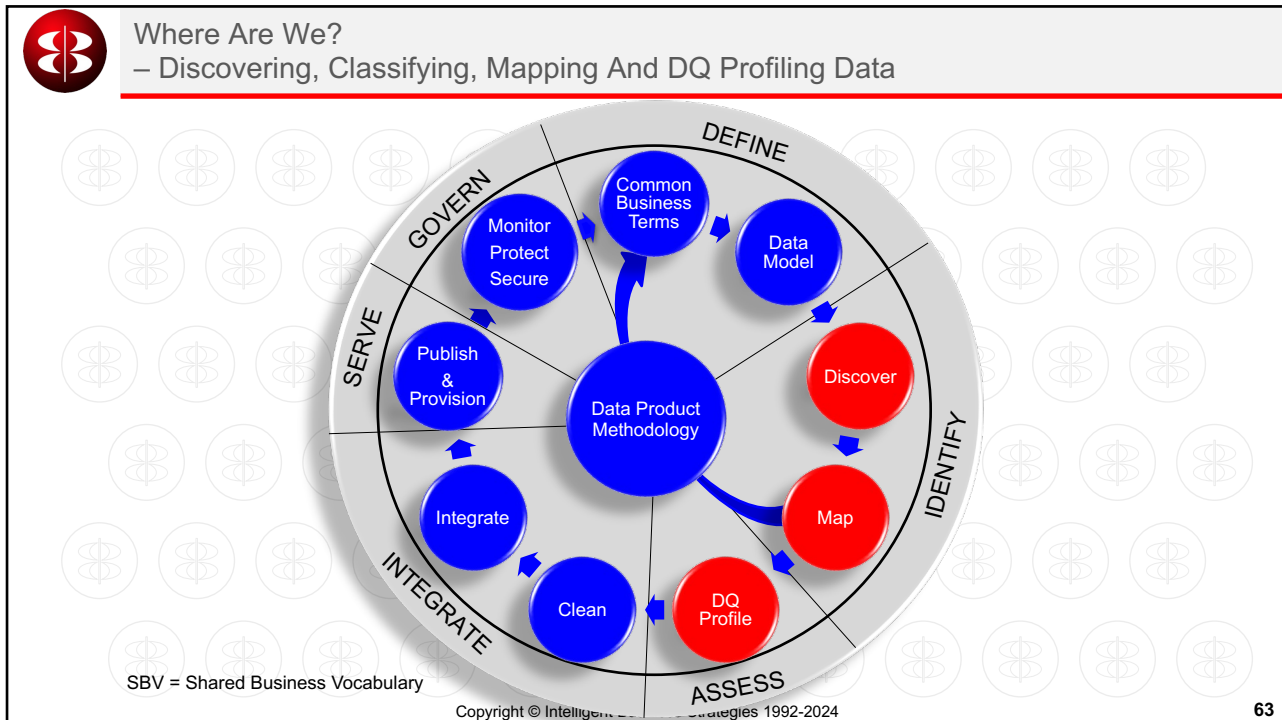


## Using Kimball As A Data Modelling Technique For Data Products – Each Dimension And Fact Data Product Is A Semantically Linked Schema Component

- Advantages
  - Create dimension and fact-based data model components separately with enterprise-wide IDs
  - Data products are data warehouse / data mart ready
  - Very easy to understand / business friendly for citizen data engineers
  - Rapid assembly of data products to create star schemas for self-service analytics
    - Fast incremental build of marts and warehouse
  - Conformed dimensions
  - Easy to aggregate
- Disadvantages
  - Data products built for analytical use only
    - Denormalised which causes data redundancy
  - Not a good choice for master data products you intended to use in business operations

Kimball





**Why Have Automated Data Discovery, Classification, DQ Profiling And Cataloguing?**

- To catalog of your data estate at scale to know what data you have
- To map your physical data assets to your glossary to understand business meaning and redundancy of data across your data estate
  - To find data across your data estate via your business glossary terms
  - To understand data origin, and know if data usage is compliant
  - To manage compliance by knowing your exposure to risk
  - To be able to control access and use
  - To know what is required by a regulation
  - To provide insights into data across your data estate
- To continually understand the quality of data in your data estate
- To classify data to enable you to protect it and govern its lifecycle
  - Identify all types of sensitive data no matter where that data resides
  - Label data according to your own data handling classification scheme
  - Label data according to your own data retention classification scheme

Copyright © Intelligent Business Strategies 1992-2024

64





## Data Catalog Product Examples – A Very Crowded Market

- Alation
- Alex Solutions
- Alteryx Connect
- Amazon Glue Catalog
- Apache Atlas (open source)
- Ataccama ONE Data Catalog
- Atlan Catalog
- BigID Data Catalog
- Boomi AtomSphere Data Platform Catalog
- Cambridge Semantic Anzo Catalog
- Cloudera Data Platform SDX Catalog
- Collibra Catalog
- Databricks Unity Catalog
- data.world
- Denodo Catalog
- Google Cloud Data Catalog
- Hitachi Vantara Lumada Data Catalog
- IBM Watson Knowledge Catalog
- Informatica Enterprise Data Catalog
- Microsoft Purview
- Oracle Cloud Infrastructure Data Catalog
- Qlik
  - [Enterprise Data Catalog](#)
  - [Qlik \(Talend\) Data Catalog](#)
- Quest (formerly erwin) Data Catalog
- Rocket Software (formerly ASG) Intelligent Data Catalog
- SAP Data Intelligence Catalog (cloud)
- SAS Information Catalog
- Salesforce Tableau Catalog
- TIBCO Cloud Metadata Catalog
- Top Quadrant TopBraid EDG Data Catalog
- Truist Zalani Arena Data Catalog

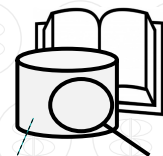
Copyright © Intelligent Business Strategies 1992-2024

65



## AI Is Being Used In Data Catalogs To Automate Data Discovery And Flag Data And Data Processing That Needs Governed

- Automated data and data relationship discovery
  - [Can support automatic data discovery on structured, semi-structured and unstructured data](#)
- Automated inference of logical entities across a distributed data landscape
- Automated lineage detection
- Automated detection of PII
  - [Pre-trained models shipped with a data catalog to automate data classification](#)
- Automated inference of non-compliance from data processing
- You can link trained models to terms in a business glossary to automate mapping



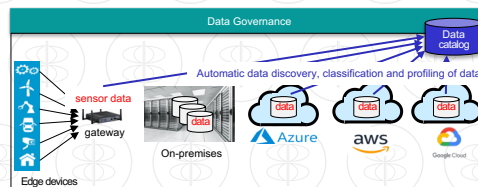
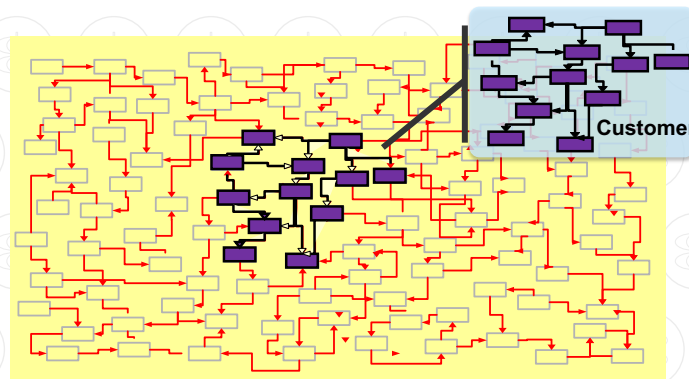
Files  
 NoSQL DBMSs  
 Cloud Storage  
 Hadoop systems  
 IoT Schema registries  
 Relational DBMSs  
 Content management systems  
 Email servers  
 ....

Copyright © Intelligent Business Strategies 1992-2024

66



Automated Data Discovery Seeks To Determine What Data You Have AND Find Complete Data Objects Across Heterogeneous Systems, E.g. Customers



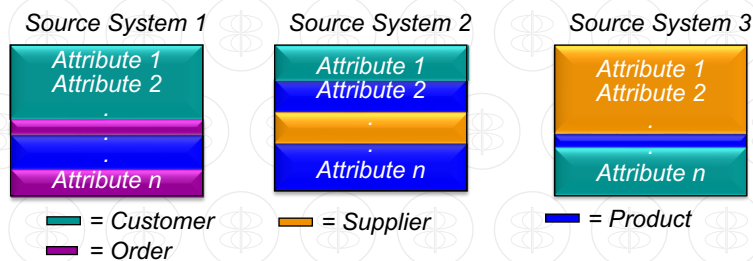
Copyright © Intelligent Business Strategies 1992-2024

67



Automatic Data Discovery Identifies Where Data Actually Exists In Disparate Systems And How It Maps To Business Data Entities In Your Business Glossary

- Identify and categorise source data by each master data entity
- Do the same for transaction data
- This makes mapping easier and enforces rigor based on the Shared Business Vocabulary



Copyright © Intelligent Business Strategies 1992-2024

68

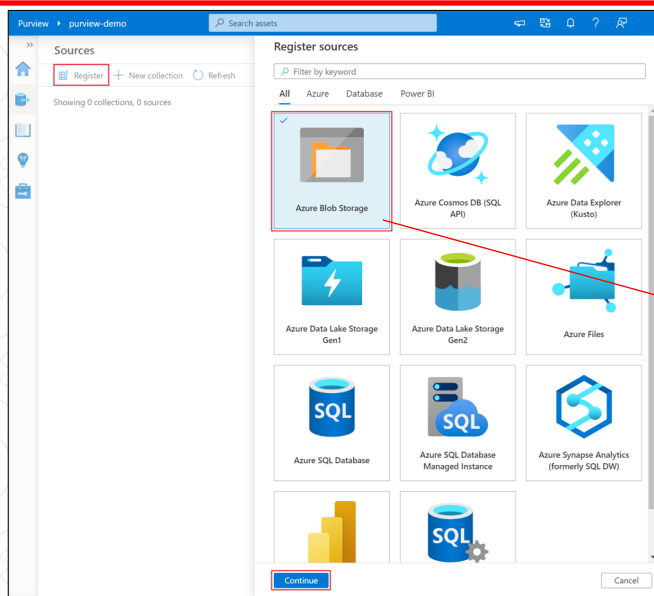


## The Data Discovery Process

- **Register** each data source
- **Set-up scanners** to discovery data on one or more data stores
- Run scanners to auto **discover, profile, classify, map** and **catalog** data in each registered data source
- **Notify** domain subject matter experts to appraise newly discovered data sources
- Collaborative **appraisal** of a newly discovered data source
- Collaborative **approval** of a newly discovered data source



## Registering Data Sources In A Data Catalog For Automated Discovery – Product Example: Microsoft Purview



Data sources can also be grouped into collections for easy management

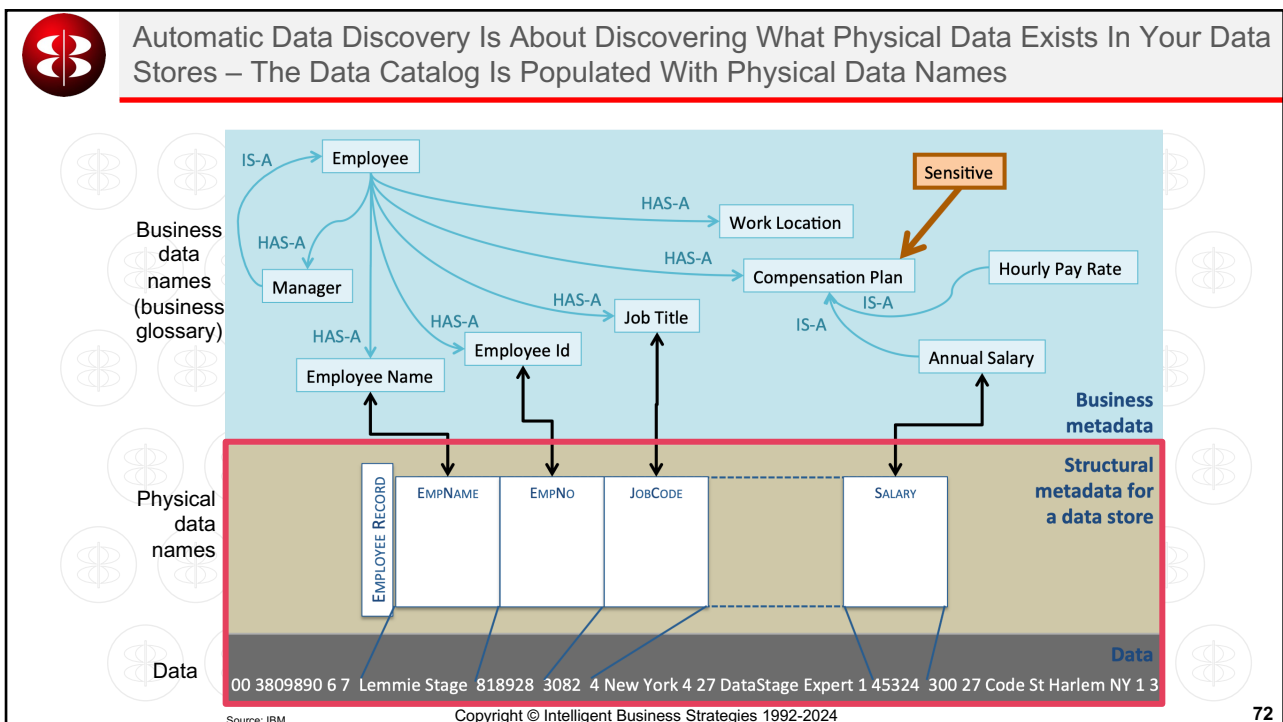
Register sources (Azure Blob Storage)

Name *	<input type="text" value="blob-storage"/>
Azure subscription	< your subscription >
Storage account name *	< your storage account >
Endpoint	<input type="text"/>
Select a collection	None

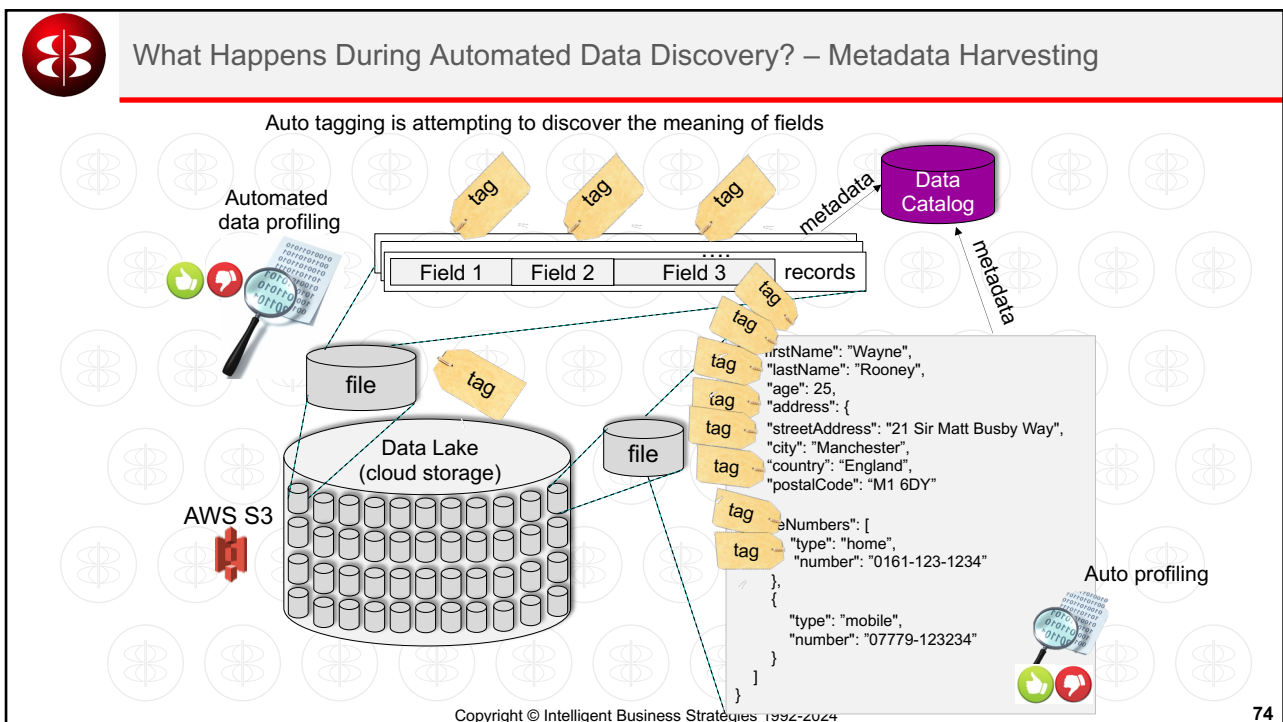
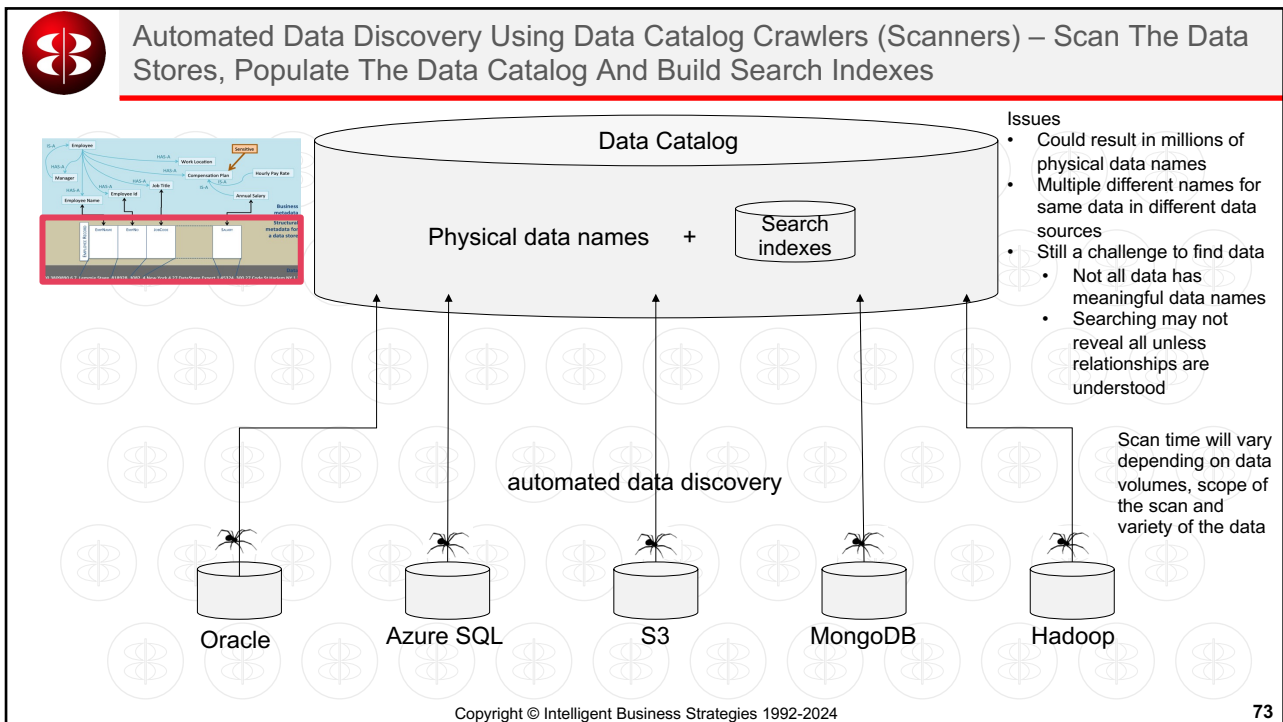
## Microsoft Purview Data Catalog – Scope The Automatic Data Discovery Scan Of The Registered Data Source

The screenshot displays the Microsoft Purview Data Catalog interface. On the left, a grid of data sources is shown, including OnPremSQLServer-Fina..., SAP-S4HANA-Procure..., AzureDataLakeStorage, Teradata-FinanceData, SAP-ECC-SalesData, AzureBlobStorage, HiveMetastore, SAP-S4HANA, AzureSQLDB-SalesIn..., FinanceSQLServer, SAP-ECC, RevenuePBIDashboa..., Teradata, and OnPremSQLServer. On the right, the 'Scope your scan' panel is active, showing a search bar and a list of assets to be scanned. The assets include AzureDataLakeStorage, analytics, productiondata, and a series of numbered folders (000-009), 00Zerotmp0, and 00Zerotmp1. A 'Continue' button is visible at the bottom of the scan configuration panel.

Source: Microsoft  
Copyright © Intelligent Business Strategies 1992-2024









## Challenges With Automatic Data Discovery What Do You Do If Your Data Catalog Cannot Connect To A Data Source?

- Data Catalogs **do not cover all data sources**
  - Data catalogs often only support a small set of packages applications if at all
  - Data catalogs often do not have connectors for many different SaaS applications
  - Also, several SaaS apps have no metadata APIs and SaaS app vendors do not allow access to their underlying DBMS
- Other niche products can be used to discovery metadata
  - Silwood Technology Safyr
    - Packaged application metadata discovery, e.g. SAP, Oracle EBS, Salesforce
    - Safyr can harvest metadata and populate several 3<sup>rd</sup> Party data catalogs
  - Manta
  - Octopai
  - Many vendor software offerings use connectors from Meta Integration Technology Inc (MITI) to get at metadata
- A popular workaround is to unload SaaS application data to files and scan the files

Copyright © Intelligent Business Strategies 1992-2024

75

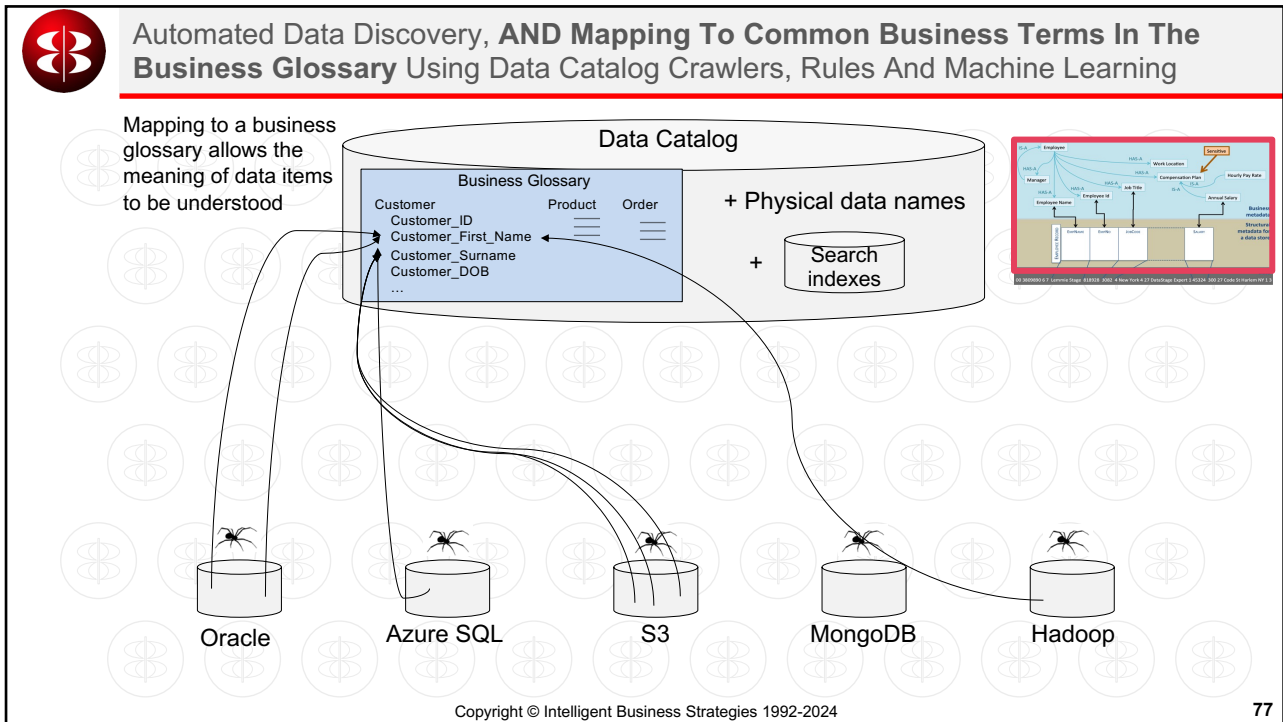


## How Do You Map Physical Data Names To Business Terms In A Business Glossary?

- Manual mapping done by business domain subject matter experts and data stewards
- Automatically using several techniques (preferred)
  - Pre-built machine learning models (algorithmic scoring)
  - AI assisted mapping
    - Auto data discovery
    - Auto data clustering of similar data
    - Observability of manual mapping to glossary terms
    - Auto labelling of similar data based on observations
    - Training machine learning models to do it automatically
  - User defined rules, e.g., regular expressions
  - Reference data

Copyright © Intelligent Business Strategies 1992-2024

76



**Enabling Glossary Business Term Association In Informatica IDMC Data Governance and Catalog To Map Physical Data Names To Business Terms in a Business Glossary**

Informatica, can automatically map physical data names to business terms in a business glossary. It does this by using an algorithm (based on accepted business terms on data domains, column similarity, and name match between a column and business term) to calculate a confidence score that it can map it correctly. You can specify a threshold for the confidence score above which it will automatically assign a business term to a physical data asset.

Source: Informatica

Copyright © Intelligent Business Strategies 1992-2024 78

## Using Regex Rules In Hitachi Vantara Lumada Data Catalog To Identify Personal Social Security Number And Map Them Your Business Glossary

The screenshot shows the 'Glossary' management interface. On the left, a tree view shows 'User-defined Tags' including 'Address', 'Food Service', 'Restaurant ID', 'Restaurant Name', 'Cuisine', and 'Inspections'. A callout box points to 'Import or create business glossary'. Another callout points to 'Manage tags'. On the right, the configuration for the 'Social Security Number' tag is shown, including a 'Tagging Rule' with a regex: `^(?!000)(?!666)(?!9)[0-9]{3}[-]?(?!00)[0-9]{2}[-]?(?!0000)[0-9]{4}$`. It also shows 'Min Length' and 'Max Length' set to 11, and 'Association Accuracy' set to 20%.

Source: Hitachi Vantara

Copyright © Intelligent Business Strategies 1992-2024

79

## Auto Discovery Of Composite Domains Such As Customer, Address, Product In Informatica Data Catalog

The screenshot displays a table named 'Order' with columns Field0 through Field8. Composite domains are overlaid on the data: 'Date' covers Field0, 'Customer' covers Field1 and Field2, 'Address' covers Field3, Field4, and Field5, 'Product' covers Field6, Field7, and Field8. A 'ProductID' domain is also shown over Field6. A callout box at the bottom shows 'Composite Data Domains' with 'Customer' and 'Order' selected, and a link to 'View Instances'.

Source: Informatica

80

## Informatica Enterprise Data Catalog – Creating A Composite Domain In Informatica Enterprise Data Catalog

**New Composite Data Domain**

Specify the name, description, and rules for the new Composite Data domain.

**General**

Name:

Description:

**Specify Match criteria**

Contains data domains

- Street x Country x
- OR
- ZipCode x
- OR
- State x

Source: Informatica

Copyright © Intelligent Business Strategies 1992-2024 81

## Selecting Composite Domain Discovery In Informatica Enterprise Data Catalog And See The Number Of Composite Domains Inferred On A Scan

**Composite Domain Discovery**

Enable Composite Domain Discovery

Composite Data Domain Settings

Select Composite Data Domain:  All Composite Data Domains  Specific Composite Data Domains

**Data Discovery**

Enable Data Discovery

Discovery Types

Discover:  Unique Key Inference  Profiling

Domain Connection Settings

Custom  Global  Select the choice

Domain Name:

Source: Informatica

Type	Schedule	Triggered By	Status
Metadata Load	----	Manual	Completed
<b>Composite Domain Inference</b>	----	Manual	Completed
Profile Run	----	Manual	Completed
Profile Result Fetch	----	Manual	Completed
Similarity Profile and Value Fr...	----	Manual	Completed
Similarity Profile and Value Fr...	----	Manual	Completed

**Logs**

Name	Value
Composite Domain Config Count	7
Composite Domain Inference Complete	Wed Apr 14 03:29:24 IST 2021
Composite Domain Inference Start	Wed Apr 14 03:26:15 IST 2021
Composite Domains to be Added to catalog	0
Composite Domains to be inferred Count	7
Composite Domains to be removed from catalog	0

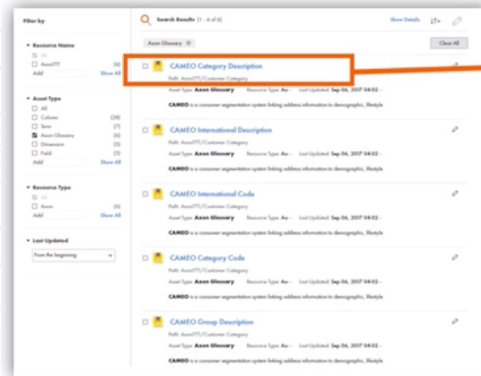
Source: Informatica

Copyright © Intelligent Business Strategies 1992-2024 82



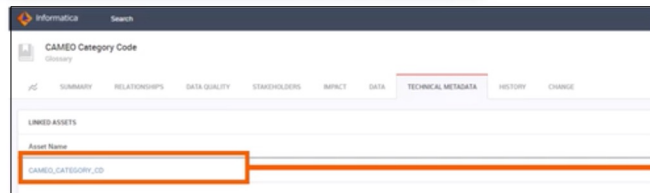


## Linking The Glossary To The Data In The Catalog – Informatica Axon Business Glossary Can Be Scanned By Enterprise Data Catalog



Business Glossaries from Informatica Axon

You can import (scan) Axon business glossaries into the catalog



Allows you to view the data catalog entries linked to the common terms in the glossary

Links to EIC from Axon



## Communities Can Then Collaborate Over Tagged Data, Changing Tags To Create A Vocabulary

Auto-tagging and profiling of discovered data followed by collaborative tag alterations and approval



“vocabulary crowd sourcing”

Create the business glossary when everyone is in agreement

Discuss, Edit, Approve, Publish



## Alation Auto Tags Discovered Data Using Machine Learning, Highlights 'Guesses' And Requests Confirmation/Rejection

**order** Columns

	Name	Title	Type	
1	id	ID	INT	
2	ordr_tp	Order Type	INT	7 distinct values
3	ordr_dt	Order Date	DATE TIME	
4	byr_id	Buyer ID	INT	
5	item_id	Item ID	INT	
6	qty	Quantity	INT	18 distinct values
7	shpg_dt	Ship Date	DATE TIME	
8	dlvry_dt	Delivery Date		
9	shpg_addr_id	Shipping Address		

Alation has guessed that shpg\_dt means "Ship Date". Is that right?

Copyright © Intelligent Business Strategies 1992-2024

85



## Alation Generated Business Terms Can Be Added To A Glossary As New Terms

**Glossaries**

Discover popular terms and add them to Glossaries to share your knowledge

Suggested Terms

Term	Abbreviations	Related Objects	Popularity	Report	Add...
customer	cust, customer, cstmr	58		Report	Add...
provider		127		Report	Add...
type	typ, type	277		Report	Add...
demographic	dem, demographic	17		Report	Add...
dem		17		Report	Add...
demo	demo, dem	17		Report	Add...
payment	paymt, payment, pymnt	93		Report	Add...
state	st, state, stat	105		Report	Add...
pay	ps, pay	142		Report	Add...
segment	seg, segment, sgmnt	47		Report	Add...
value		47		Report	Add...
date	dt, date	143		Report	Add...
rate	rat, rt	75		Report	Add...
total	tot, tot, t	83		Report	Add...
number	num, number	93		Report	Add...
group	grp, group	52		Report	Add...
diagnosis related group	diag	23		Report	Add...
Diagnosis Related Group	Diagnosis Related Group, diag	23		Report	Add...
word		13		Report	Add...
source	source, src, sr	115		Report	Add...

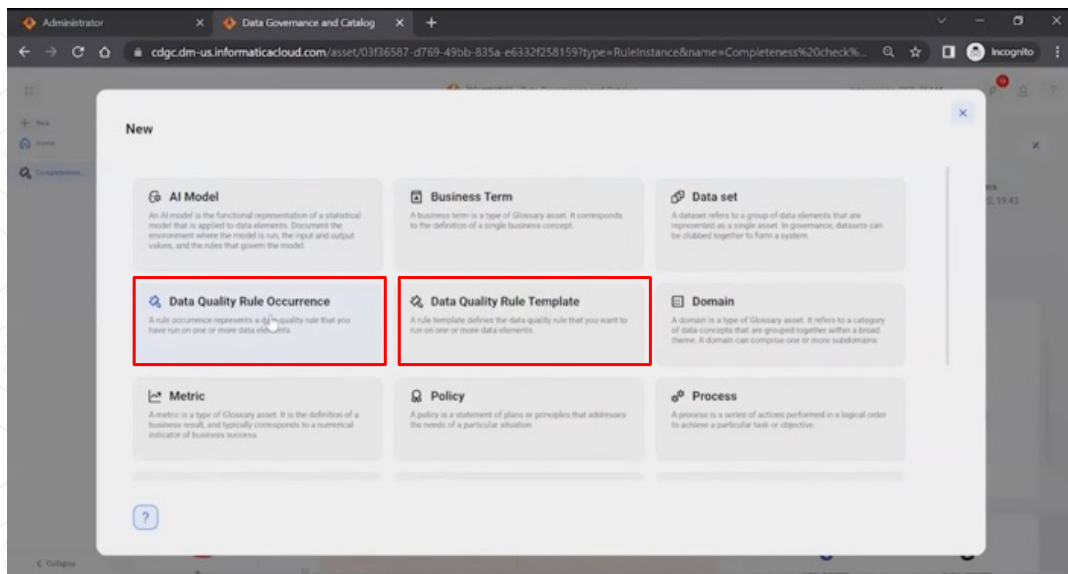
As New Glossary Term Add this suggested term and its related objects to a selected Glossary as a new Article.

Copyright © Intelligent Business Strategies 1992-2024

86



## Data Quality Rules Can Be Defined For Data Quality Profiling To Automatically Validate Data Quality During A Scan – E.g., Informatica IDMC Data Governance And Catalog



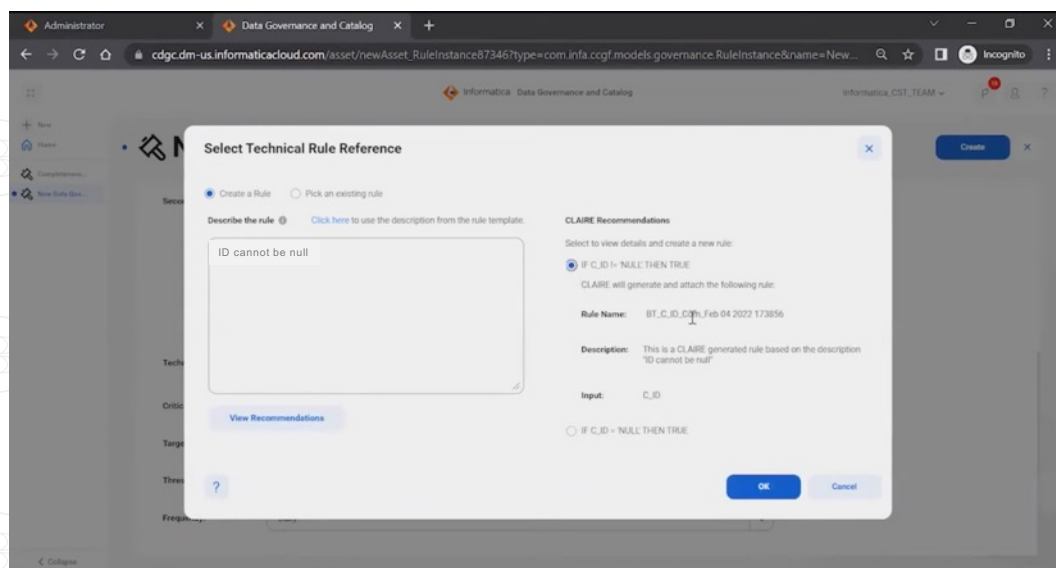
Source: Informatica

Copyright © Intelligent Business Strategies 1992-2024

87



## Informatica Allows Data Quality Rules To Be Defined Using Natural Language And Its CLAIRE AI Engine Will Recommend A Data Quality Rule



Source: Informatica

Copyright © Intelligent Business Strategies 1992-2024

88



## Data Catalogs Can Perform Automatic Data Quality Profiling And Generate Metadata In The Data Catalog – E.g., Ataccama ONE Data Catalog

Source: Ataccama

Copyright © Intelligent Business Strategies 1992-2024



## Data Quality Evaluation Rules And Cleansing Policies Can Be Assigned At The Glossary Term Level So That Any Data Mapped To The Term Is Auto Cleaned

Ataccama ONE can apply policies and rules at the glossary level to recognise data, validate it and clean it

Source: Ataccama

Copyright © Intelligent Business Strategies 1992-2024



## Automatic Data Quality Profiling To Determine Data Quality And Cataloging It – IBM Cloud Pak For Data And Watson Knowledge Catalog

Quickly profile the data source for an overview of data quality using sampling – a good option for large data sets

**Quick scan**  
Choose quick scan if you want to quickly get a general overview of the quality of your data based on the analysis of data sample. The impact of source issues is skipped. Quick scan is suitable for large data sources.

**Automated discovery**  
Choose automated discovery if you want to see the details about the quality of your data based on an in-depth analysis of all assets. The source assets are imported. Automated discovery is suitable for smaller data sources, or selected components of larger data sources.

This is much longer running as it does a full analysis of your data to provide detailed data quality insights

Source: IBM

Discovery options:  
 Analyze columns  
 Analyze data quality  
 Assign terms  
 Use machine learning to assign terms  
 Use data sampling  
The maximum number of records included in the data set sample: 1000

Source: IBM

Automated full and incremental data quality profiling is also supported by other vendors, e.g. Informatica EDC



## IBM Cloud Pak For Data Automated Data Quality Profiling – Data Quality Dashboard Of A Specific Data Source

Dashboard Data assets Data rules Relationships Column similarity Settings

**DataLakeWorkspace**

Description: Workspace with optimal settings to run a quick ana... Show more

Data assets: 6  
PII data assets: 0/6  
Reviewed data assets: 0/6  
Connections: 1  
Critical data issues: 0  
Created by: isadmin  
Last modified: Mar 23, 2020

Governed Sampling enabled

**Data quality threshold**  
6 Confirmed 100%

**Quality score distribution**

**Analysis**

**Relationships**

**Rule run status**  
3 Confirmed 100%

**Top 5 selected classes**





## IBM Cloud Pak For Data Automated Data Quality Profiling Highlights Data Quality Issues, E.g., Data Violates The Matching Rules Set Up In A Data Class

**Data quality score** 78% 0%

**Column data quality dimension results (11)**

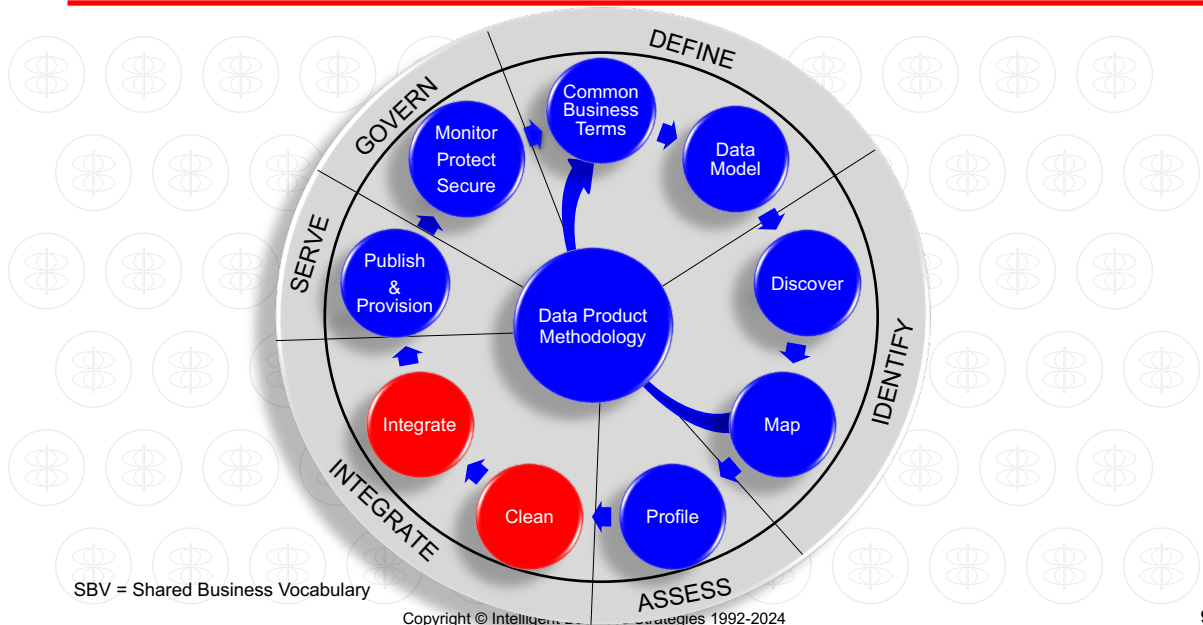
Dimension name	# of findings	% of findings	Delta	Ignore
Data class violations	106	21%	--	<input type="checkbox"/>
Inconsistent capitalization	3	1%	--	<input type="checkbox"/>
Data type violations	0	0%	--	<input type="checkbox"/>
Duplicated values	0	0%	--	<input type="checkbox"/>
Format violations	0	0%	--	<input type="checkbox"/>
Inconsistent representation of missing values	0	0%	--	<input type="checkbox"/>
Missing values	0	0%	--	<input type="checkbox"/>
Suspect values	0	0%	--	<input type="checkbox"/>
Suspect values in correlated columns	0	0%	--	<input type="checkbox"/>
Values out of range	0	0%	--	<input type="checkbox"/>

Source: IBM

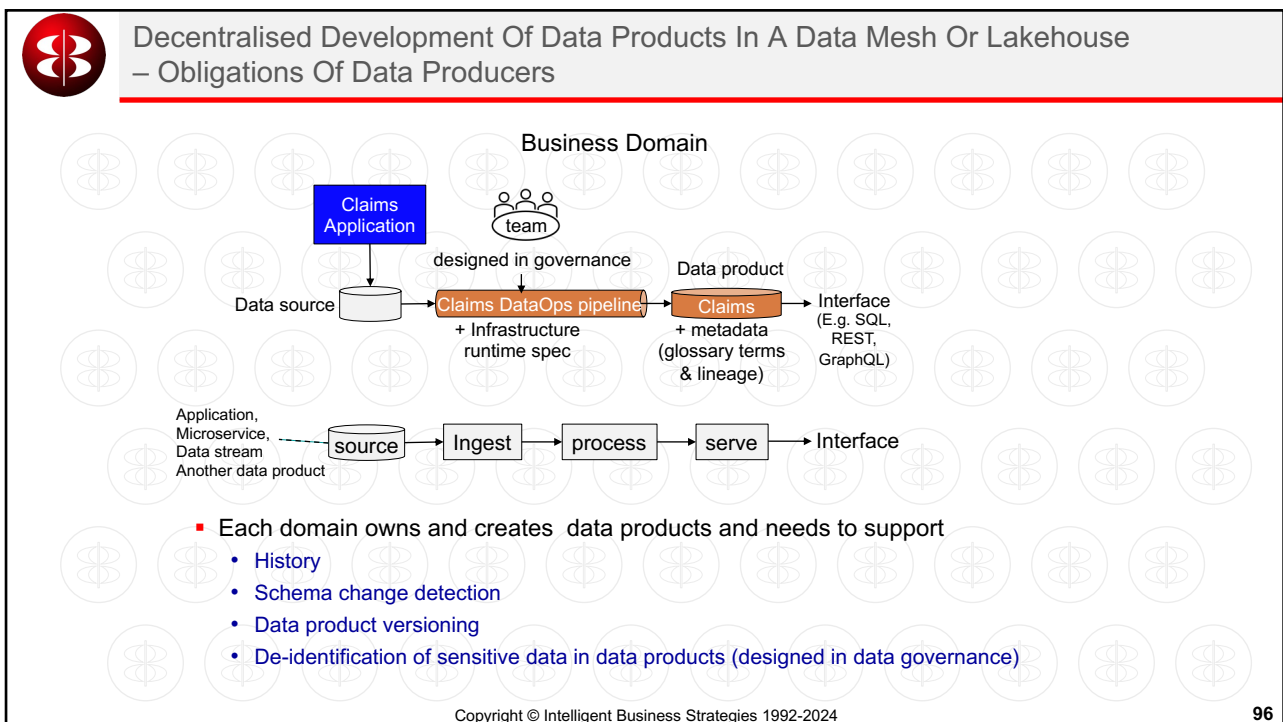
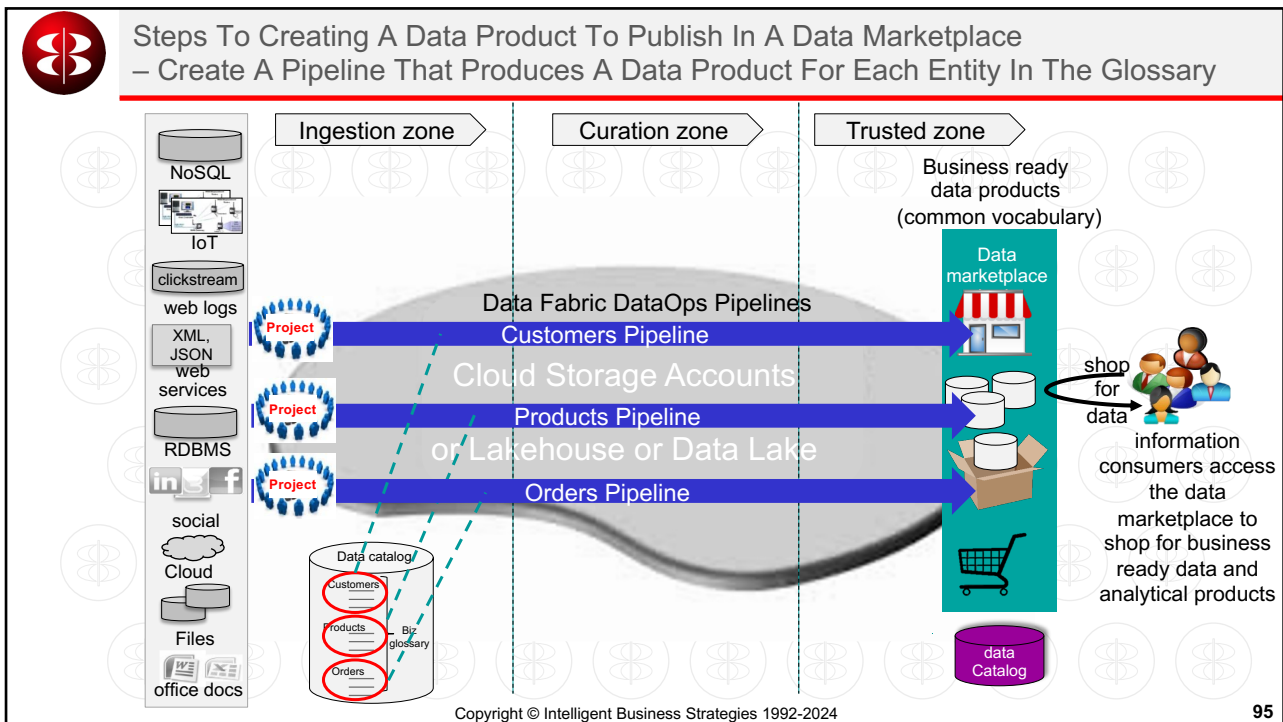
93

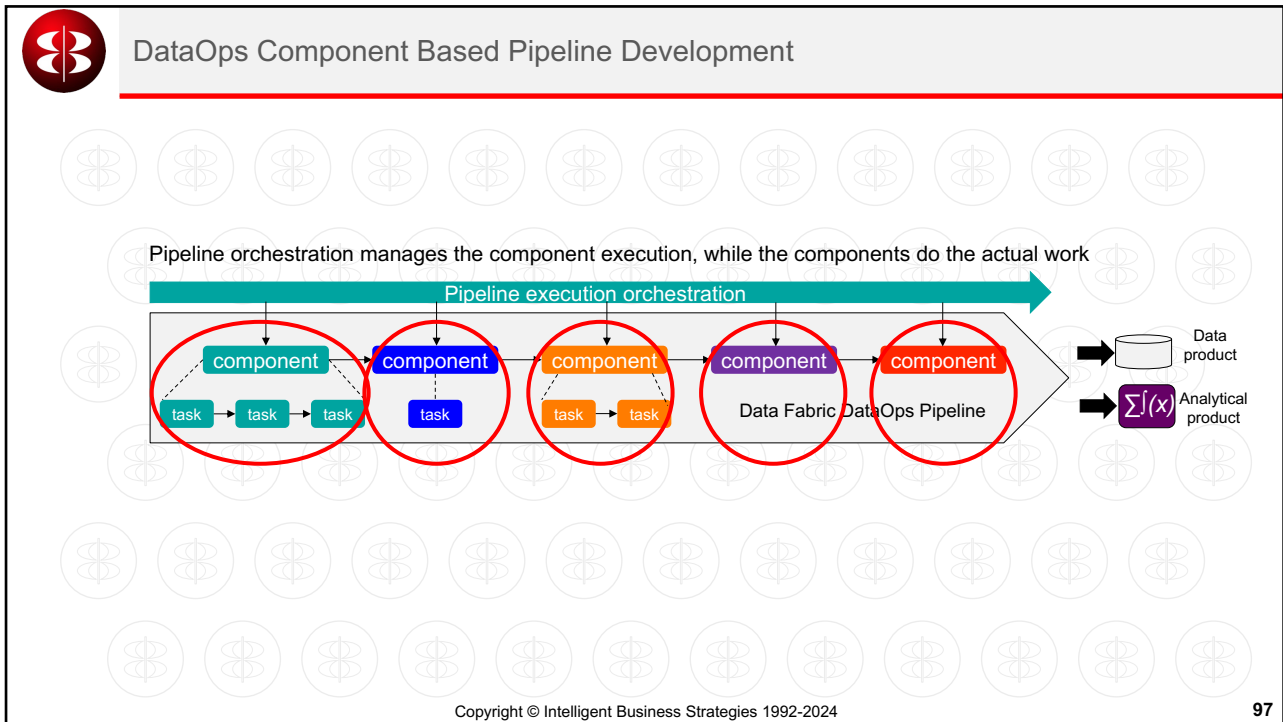


## Where Are We? – Building Data Engineering Pipelines To Integrate Data To Produce Data Products



94





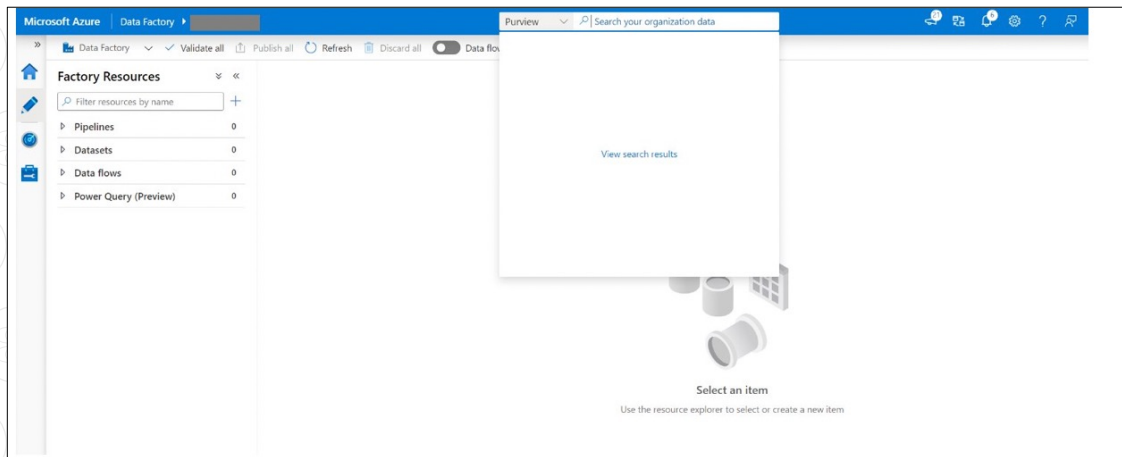
Type of Component	Examples
Data connectivity components	<ul style="list-style-type: none"> <li>Connectors to multiple data sources</li> </ul>
Data ingestion components	<ul style="list-style-type: none"> <li>File ingestion</li> <li>Database table ingestion service</li> <li>Stream ingestion service</li> </ul>
Data transportation components	<ul style="list-style-type: none"> <li>FTP service</li> </ul>
Data governance components	<ul style="list-style-type: none"> <li>Data validation services</li> <li>Data cleansing services                             <ul style="list-style-type: none"> <li>e.g. Address cleansing / enrichment</li> </ul> </li> <li>Data privacy masking / obfuscation service</li> <li>Data encryption service</li> <li>Logging and auditing services</li> </ul>
Data transformation components	<ul style="list-style-type: none"> <li>Data type transformation services</li> </ul>
Data matching and integration components	<ul style="list-style-type: none"> <li>Survivorship rule services</li> </ul>
Cognitive / ML components	<ul style="list-style-type: none"> <li>Voice-to-text conversion</li> <li>Customer segmentation clustering service</li> <li>Customer sentiment scoring model</li> <li>Customer propensity to churn scoring model</li> </ul>
Data loading components	<ul style="list-style-type: none"> <li>Bulk load service, append service</li> </ul>
Action components	<ul style="list-style-type: none"> <li>Alerts, recommendations, automation,....</li> </ul>

Reusable component libraries available in the data fabric

Copyright © Intelligent Business Strategies 1992-2024 98



## Data Product Producers Can Search The Data Catalog For Raw Data From Within Data Fabric Software – E.g. Search Microsoft Purview From Azure Data Factory



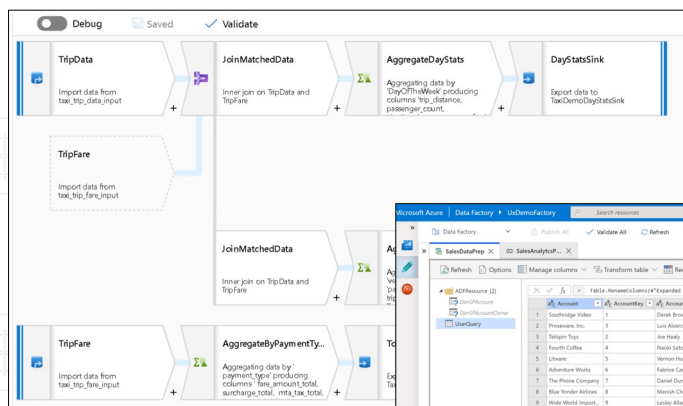
Find the data in the catalog, connect to it and start building ETL pipelines

Copyright © Intelligent Business Strategies 1992-2024

99



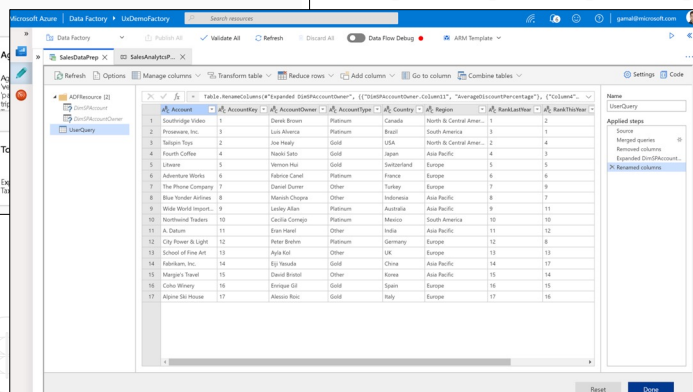
## Unified Data Integration – Azure Data Factory Enables Collaborative Business And IT Development With Mapping And PowerQuery Data Flows



Azure Data Factory Mapping Data Flows for IT professionals



Azure Data Factory PowerQuery Data Flows for business users

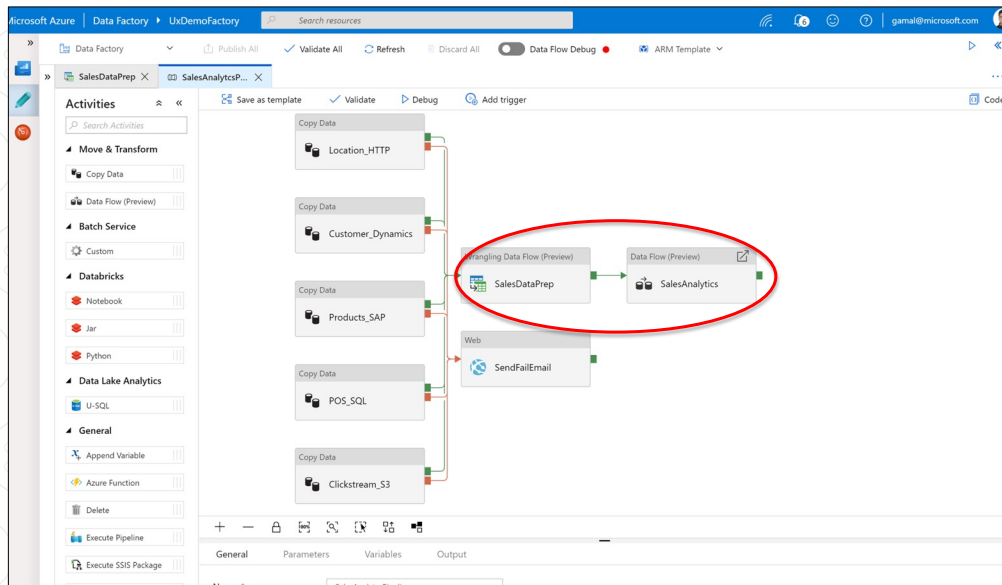


Copyright © Intelligent Business Strategies 1992-2024

100



## Azure Data Factory Unified Data Integration – Orchestrate PowerQuery And Mapping Data Flows In The Same DataOps Pipeline To Produce A Data Product

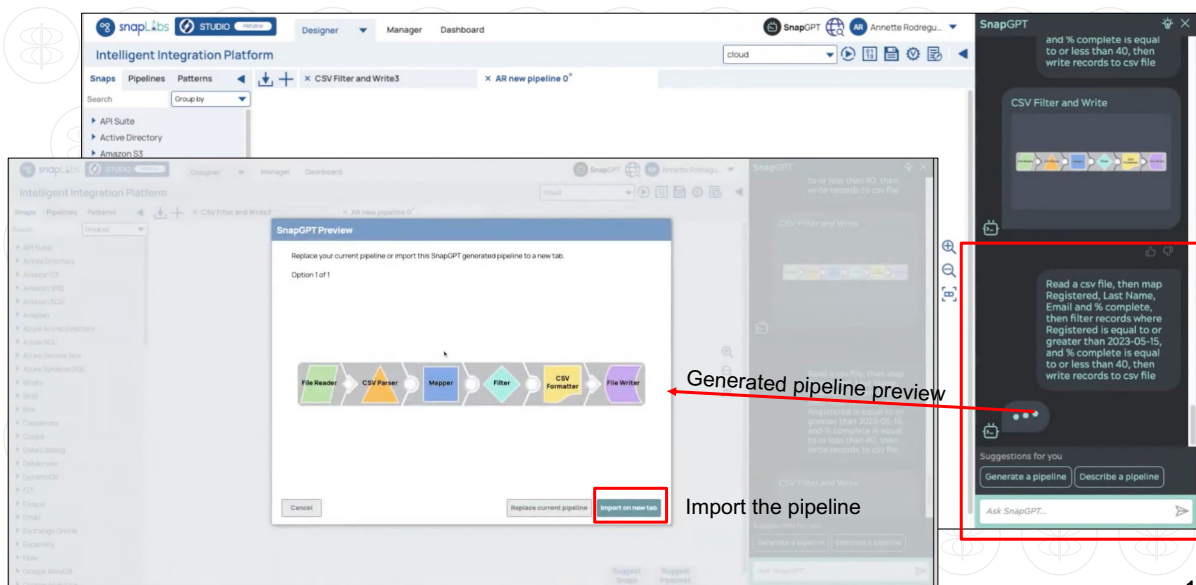


Copyright © Intelligent Business Strategies 1992-2024

101




## Generative AI In Data Engineering Opens Up Data Product Development To Citizen Data Engineers – SnapLogic SnapGPT Prompt Based Data Engineering



Copyright © Intelligent Business Strategies 1992-2024


102





## Generative AI Prompt Based Data Engineering – SnapLogic SnapGPT Pipeline Configuration Wizard


Data engineering



Source: SnapLogic

Copyright © Intelligent Business Strategies 1992-2024

103



## What If You Have To Process Text Is A Key Part Of A Pipeline?

What Is Text Analytics?– e.g. deriving structured data from unstructured content, or sentiment scoring

The proposed merger between Mega, Inc. and CNA Systems, Incorporated, has been postponed. Mega CEO Joe Smith said in an analyst call. "CNA's 1st quarter revenue dropped by 32%, and they lost 23 million dollars," Smith explained. CNA Systems sources blame weak sales in China. CNA shares (CNAI) fell 47 percent to \$9.84 on May 12, the first trading day after the announcement.

<b>Company</b>	Mega, Inc., CNA Systems, Incorporated
<b>Date</b>	May 12
<b>Person</b>	Joe Smith
<b>Person Position</b>	Mega CEO
<b>Currency</b>	23 million dollars, \$9.84
<b>Measurement</b>	32%, 47 percent
<b>Country</b>	China
<b>Concept</b>	proposed merger, analyst call, 1st quarter revenue weak sales, first trading day
<b>Event: M&amp;A</b>	The proposed merger between Mega, Inc. and CNA Systems, Inc. has been postponed

- Popular data sources include
  - Social media, email, news articles, on-line forums
- Requires pre-processing prior to analysis
  - Parsing, correction, stemming, phrase extraction, semantic grouping

Copyright © Intelligent Business Strategies 1992-2024

104

## Parsing And Extracting Entities From Text Data Using Natural Language Processing In The Pipeline – E.g. SnapLogic

Source: SnapLogic Copyright © Intelligent Business Strategies 1992-2024 105

## Built-In ML And Sentiment Analysis Now Ships Out-of-the-Box In Self-Service Data Preparation In Cloud BI Tools – E.g. Oracle Analytics Cloud Dataflow

Data Set	Columns	Treat As	Default Aggregation
Date		Attribute	
Price		Measure	Average
Market Cap		Measure	Average
Total Volume		Measure	Average

Source: Oracle Copyright © Intelligent Business Strategies 1992-2024 106

### Processing Streaming Text And Textual Data At Rest Using Cognitive Services In A Pipeline To Improving Voice Of Customer Analysis - Amazon

Source: aws

Copyright © Intelligent Business Strategies 1992-2024

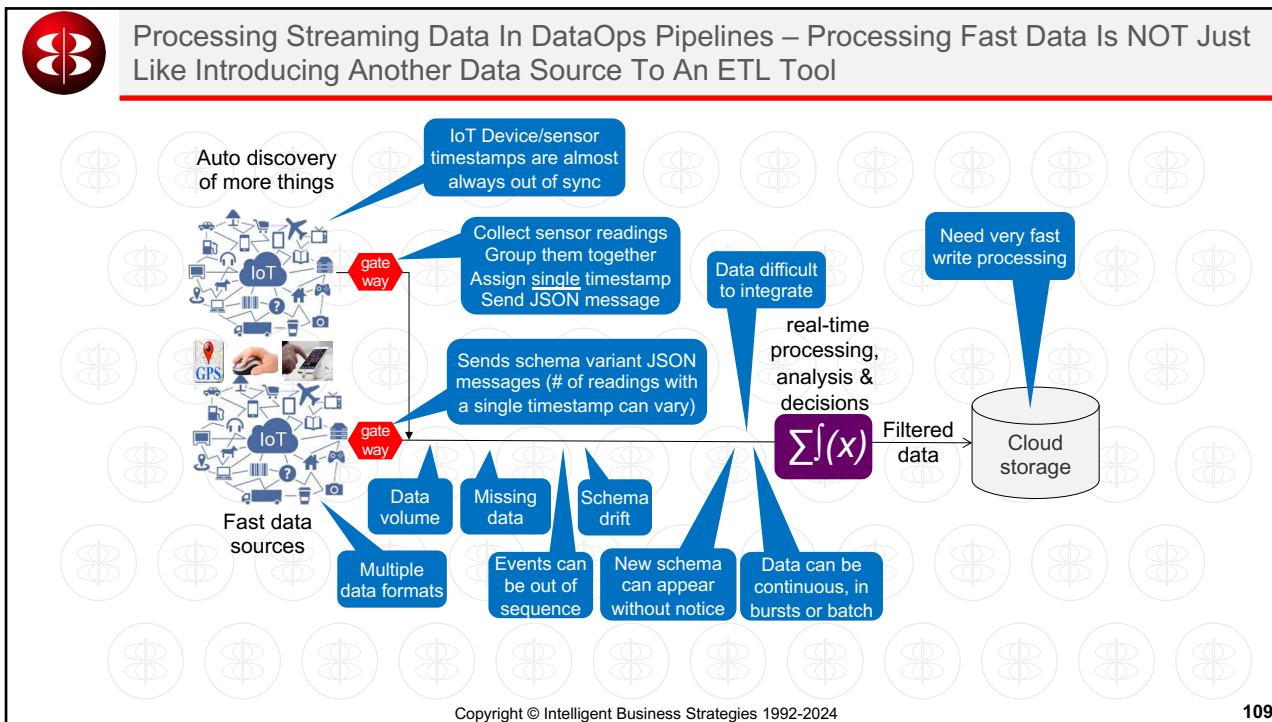
107

### Tools Are Now Appearing To Build End-To-End Data And Analytical Pipelines – E.g. Microsoft Data Factory

Launch a Databricks notebook on Spark e.g. to analyse data

Copyright © Intelligent Business Strategies 1992-2024

108



### Many Data Fabric Platforms Streaming Data Messaging Services As A Data Source – e.g. IBM Cloud Pak For Data Supports Kafka And Others' Services

IBM Cloud Pak for Data

Add connection

Create a new connection, select an existing connection from the list of platform connections, or connect to a service.

Provider

- IBM
- Third-party

Compatible services

- Catalogs
- Data Virtualization
- DataStage
- Metadata Import
- Watson Studio

Find connection types

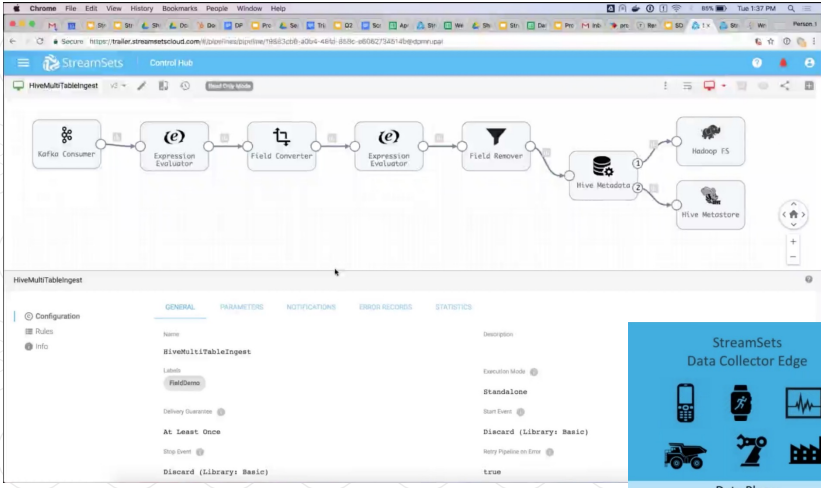
- Apache Derby
- Apache HDFS
- Apache Hive
- Apache Kafka
- Box
- Cloud Object Sto...
- Cloud Object Sto...
- Cloudant
- Db2 Big SQL
- Db2 Event Store
- Db2 for i
- Db2 for z/OS
- Db2 Hosted
- Db2 on Cloud
- Db2 Warehouse
- Dropbox
- Microsoft Azure ...
- Microsoft Azure ...
- Microsoft Azure ...
- Microsoft Azure ...
- Microsoft Azure ...
- Microsoft SQL ...
- MongoDB
- MySQL
- Salesforce.com (...)
- SAP ASE
- SAP IQ
- SAP OData
- Snowflake
- SQL Query
- Tableau
- Teradata

Connect directly to streaming data in motion

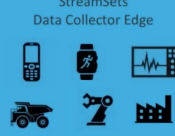
Copyright © Intelligent Business Strategies 1992-2024

110

## SoftwareAG StreamSets – Supports Data Collector Edge, Kafka Plus Automatic Schema Drift Detection And Handling



StreamSets  
Data Collector Edge



Data Plane

Micro-footprint headless agent for operation of data pipelines on edge devices

- Developers
- Scientists
- Architects

Copyright © Intelligent Business Strategies 1992-2024

111

## Schema Drift Can Be Managed In Several Tools – E.g. Microsoft Azure Data Factory Mapping Flows

Source Settings
Dataset settings
Schema
Inspect

Output stream name \*

Source Dataset \* movie\_dataflow\_source ✎ Edit + New

Options  Allow schema drift ⓘ

Sampling \*  Select Allow Schema Drift if the source columns will change often. This setting will allow all incoming fields from your source to flow through the transformations to the Sink.

Without handling for Schema Drift, your ETL jobs are likely to be impacted by in data source changes

Map all new fields in the Sink Transformation so that all new fields get picked-up and landed in your destination

Sink      Dataset settings      Mapping

Auto Map  On ⓘ

You can add transformations that can handle schema drift using pattern matching to match columns by name, type and value

Enter expression... ANY + -

Add column
Add column pattern

Clicking this in the Derived Column or Aggregate transformation allows you to create a transformation that understands "Schema Drift"

1992-2024

112



## Schema Drift In Microsoft Azure Data Factory

Azure Data Factory treats schema drift flows as late-binding flows, so when you build your transformations, the column names will not be available to you in the schema views throughout the flow

\$\$ represents each matched column from your matching pattern, e.g. append the text “\_total” to each column is a ‘double’ data type

113

## DataOps – Component Based Development Using A Common Version Control System

Each component of the pipeline is a new, independent branch. Components are merged into the main branch as they are completed

Branch and merge enables collaborative development with different people working on different components

Copyright © Intelligent Business Strategies 1992-2024

114

## Several Data Fabric Platforms Support DevOps & Continuous Integration / Continuous Deployment, E.g. Microsoft Azure Data Factory/GitHub Integration

**Option 1**  
(ADF home page)

**Option 2**  
(ADF authoring canvas)

Configuring a GitHub repository with Azure Data Factory

**Code Repository configuration**

Source: Microsoft Copyright © Intelligent Business Strategies 1992-2024 115

## Several Data Fabric Platforms Now Support DevOps With Continuous Integration / Continuous Deployment, E.g. Microsoft ADF GitHub Version Control

**1** Users can create feature branches

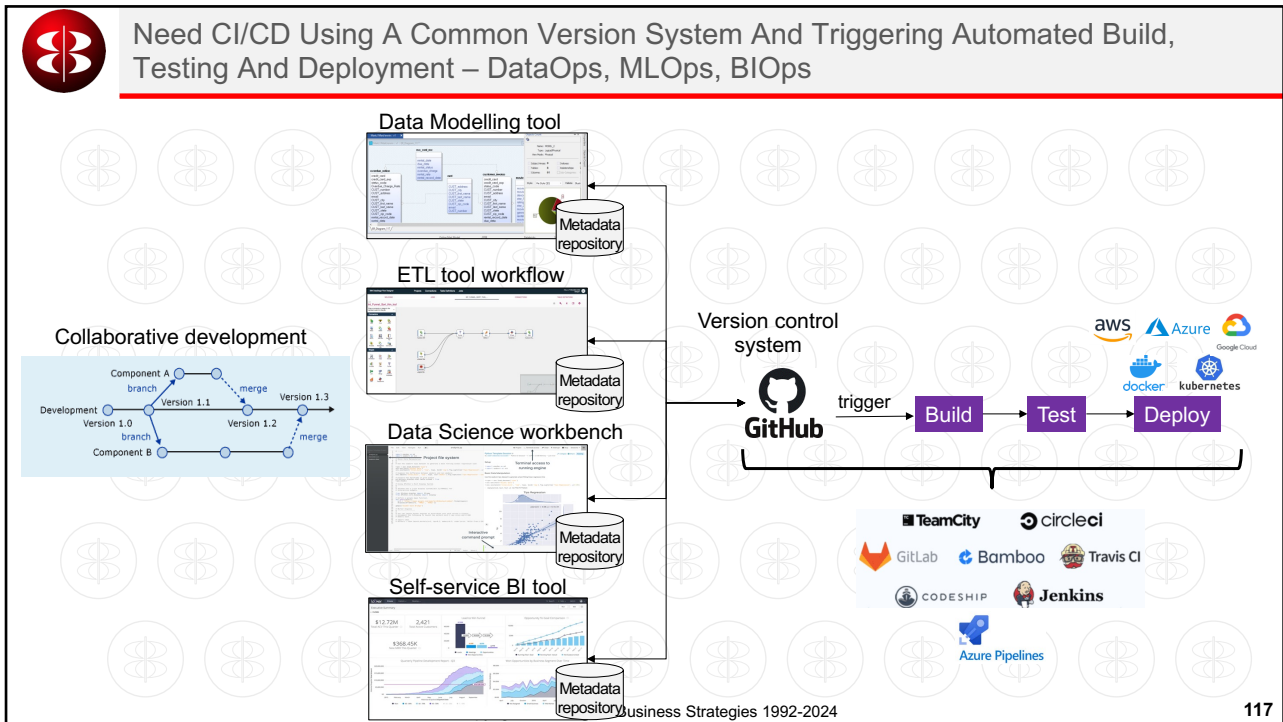
**2** Merge the changes from your feature branch to your collaboration (master) branch

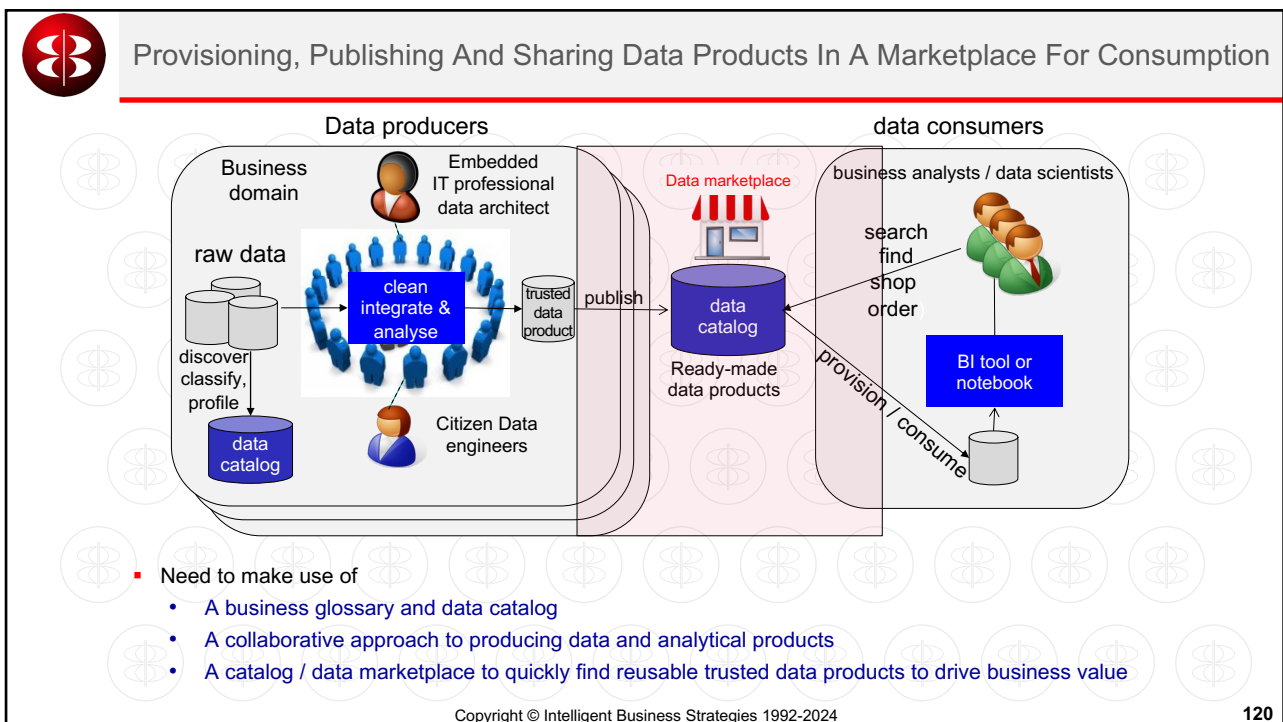
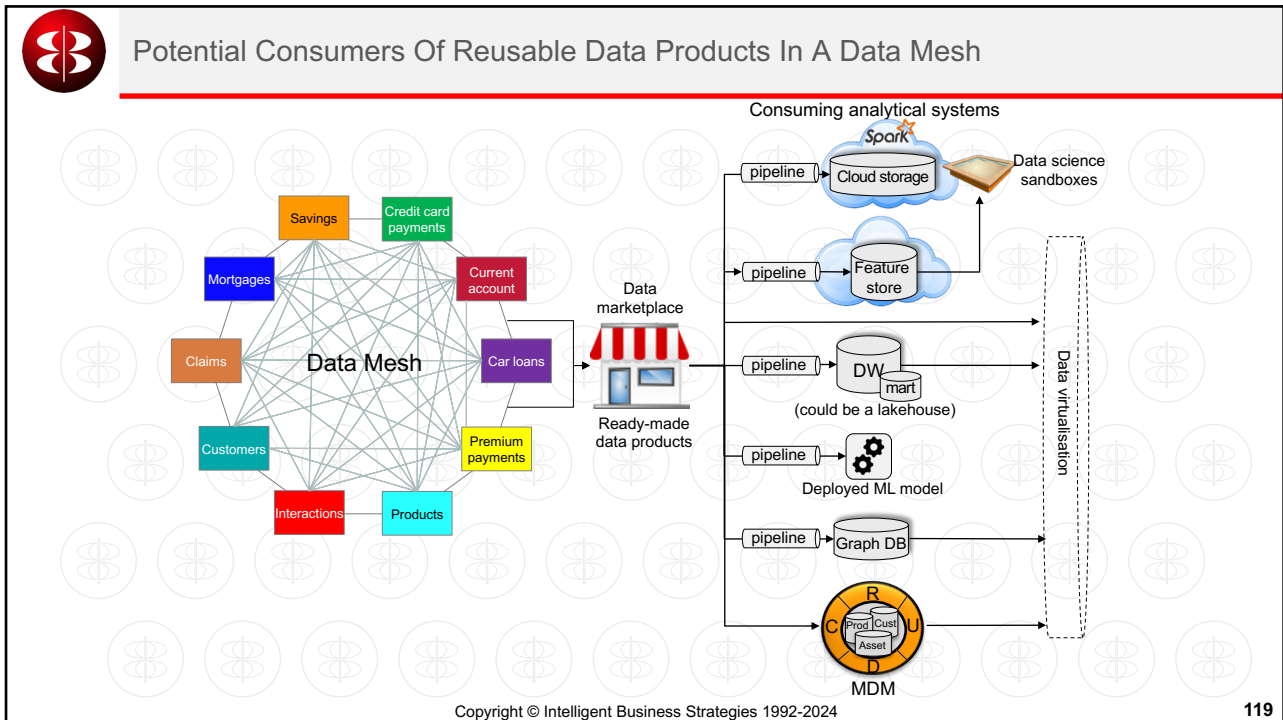
**3** Publish code changes in the master branch to the Data Factory service

Factory resources include data sets, data flows, integration run times (e.g. to get at on-premises data sources), linked services, etc.

**4** Confirm that the publish branch and pending changes are correct

Copyright © Intelligent Business Strategies 1992-2024 116









## What Is An Enterprise Data Marketplace?

### Enterprise Data Marketplace

A catalog application that governs the publishing, sharing and use of read-made, trusted, data and analytical products that are available as services with common data names documented in a business glossary, full metadata lineage. These data and analytical products are tagged and organised to make them easy to find, access, share and reuse across the enterprise

Product examples:

- Alation Marketplaces
- Amazon Datazone
- Collibra
- Databricks Data Marketplace
- Harbr Data
- Informatica IDMC Data Marketplace
- Quest Data Marketplace
- Snowflake Data Marketplace



## Trusted Business Ready Data Products In An Enterprise Data Marketplace For Users To Consume And Use

### Data available as a Service



**BUSINESS  
READY**

Business ready data products can be logical entities

#### Master Data

- Customers
- Products
- Suppliers
- Assets
- Employees
- Materials

#### Transaction Data

- Orders
- Shipments
- Payments
- Adjustments
- Returns





## Data Sharing Objectives

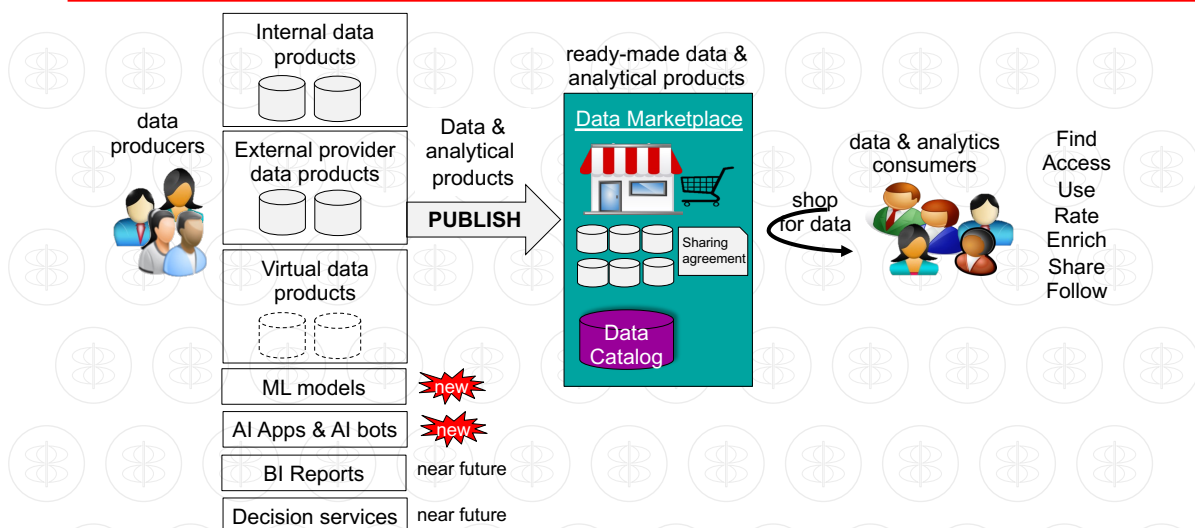
- To produce trusted, compliant, secure data products published and available for consumption
- To enable safe, secure, compliant sharing of data with data consumers
- To define terms of use of data in a data sharing agreement to ensure it is handled correctly
- To ensure the data sharing agreement is accepted before sharing data
- To monitor, track and audit
  - Email, web chat, FTP, cloud storage and other ways data can be shared
  - Who outgoing and incoming data is shared with
  - Acceptance, receipt and use of shared data
- To prevent sharing of sensitive data including all cross-border data transfers if
  - Not permitted by legislation in the jurisdiction the data originates
  - Consent is not given
- To be able to share data without moving it
- To be able to report on all data sharing activity

Copyright © Intelligent Business Strategies 1992-2024

123



## Enterprise Data Marketplace Operations To Facilitate The Publishing, Sharing And Governance Of Data And Analytical Products



Snowflake apps can be shared in Snowflake Marketplace  
 Databricks Lakehouse Apps can be shared in Databricks Marketplace

Copyright © Intelligent Business Strategies 1992-2024

124

### Key Elements In Enterprise Data Sharing – Governance Publishing Of Data Products

- Quality assurance and governance of data products is needed before they are published in a marketplace for consumption

Business Glossary + Data Lineage + Data Freshness + Data Owners	Data Quality score	✓
	Common data names (business glossary)	✓
	Full metadata lineage	✓
	Access security policy defined	✓
	Privacy policy defined	✓
	Retention policy defined	✓
	Version management	✓
	<b>Data sharing agreement defined</b>	✓
	Tagged by business objective	✓

Data product → Data marketplace

Labels: Data product, Data marketplace, Data Catalog, Data product quality assurance

Copyright © Intelligent Business Strategies 1992-2024 125

### Companies Should Create Semantically Linked Data Products – Why Do This?

Customer, Product, Orders → Business Glossary (Data catalog) → Data Producer Pipelines → Assemble data products

Data consumers → Analytical use case (DW & marts, Jupyter, GraphDB)

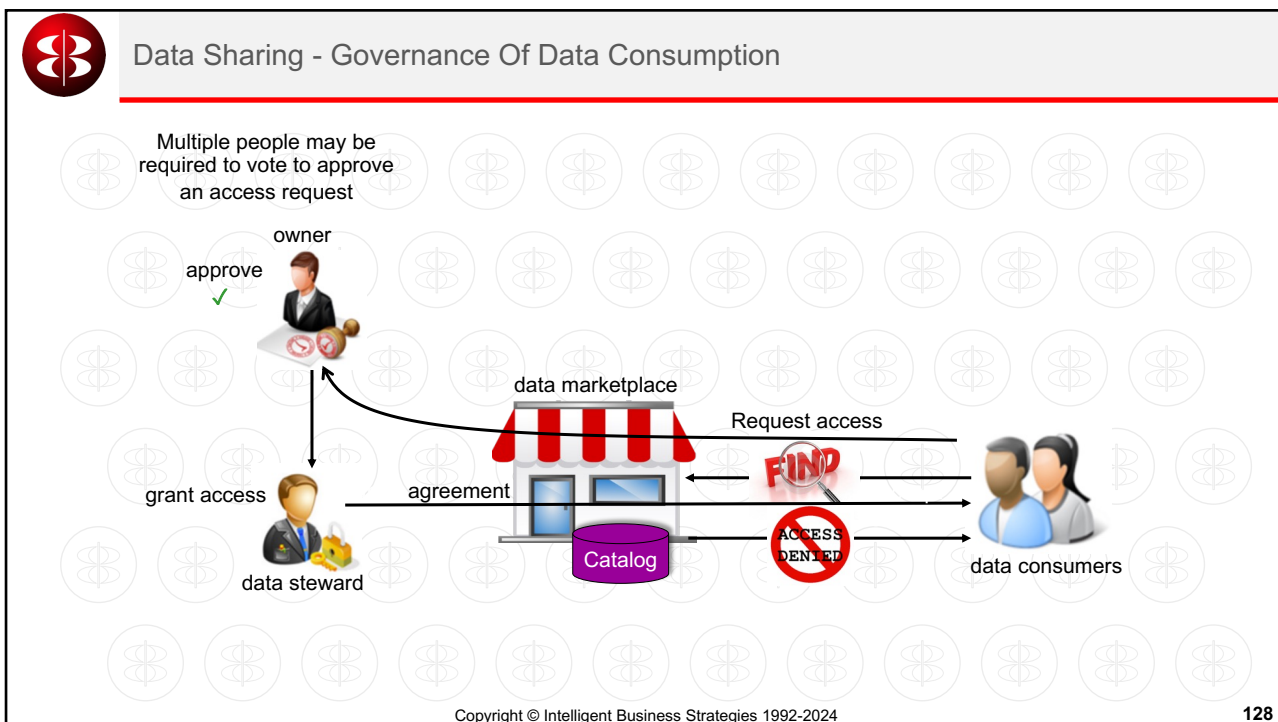
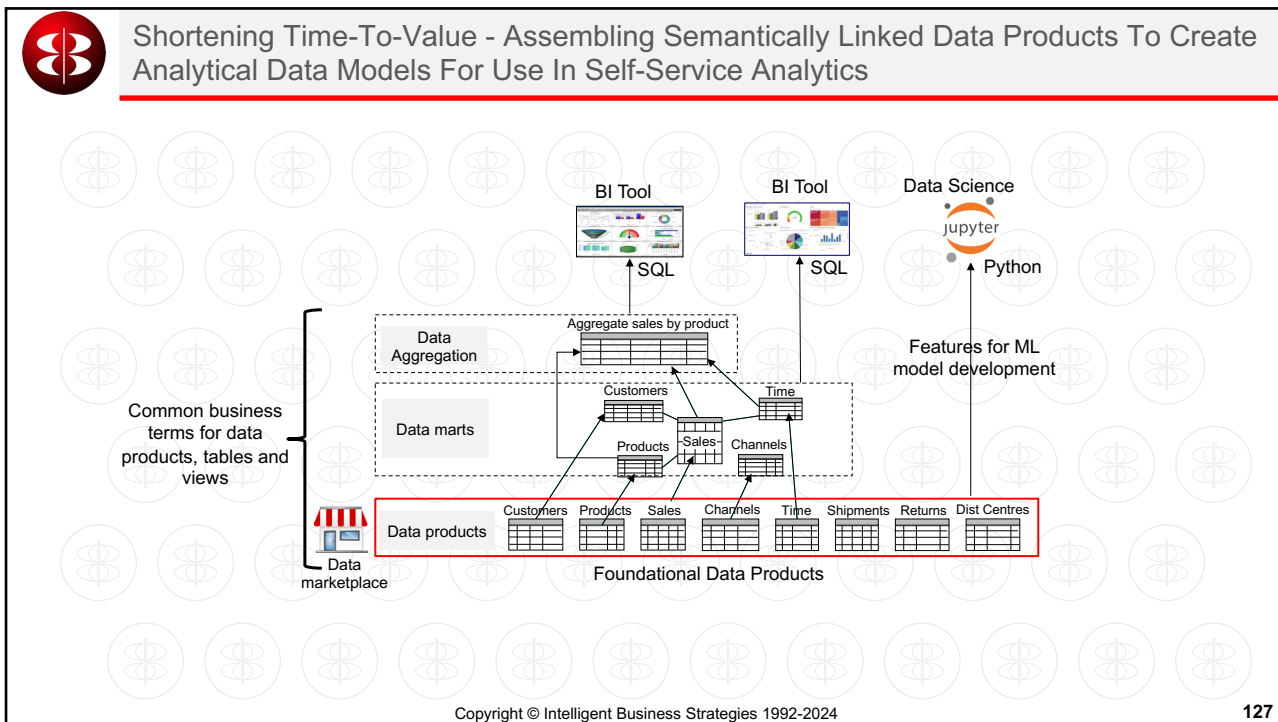
Enterprise-wide Identifiers allows quick assembly of 'component' data products for different analytical use cases

Like Lego bricks

e.g., Orders should reuse the keys of other data products such as Customer and Product

consume, assemble & use to meet analytical needs

Copyright © Intelligent Business Strategies 1992-2024 126





## Information Consumers Shopping For Data In An Enterprise Data Marketplace – E.g. Informatica Axon Data Marketplace

The screenshot shows the 'Welcome to Data Marketplace' interface. It features a search bar with the text 'Find Data Collections by name or purpose'. Below the search bar, there are several data collection cards, each with a rating (stars), a title, a description, and a progress indicator. The cards include 'Fleet DOE Product', 'Finance Reporting', 'Customer Samples', '2020 Recall Protocol', 'Staffing Roster', and 'Temp Customer Data'. On the right side, there is a 'TOP 5 ORDERS BY CATEGORY' bar chart and a 'RECENT COMMENTS' section with user feedback.

Audit history, reporting, analytics on data consumption patterns

Policy classification and privacy enforcement for sensitive data



## Informatica Data Marketplace – Shop And Checkout

The screenshot shows the 'Shop And Checkout' interface. It features a 'Data Quality Breakdown' section with a 98% score and a 'Data Sets' table. The 'Data Sets' table has columns for Name, Definition, Ref., Lifecycle, System Short Name, Attributes, and Data Quality. Below the table, there is an 'Attributes' section with a table of attributes. A 'Checkout' button is highlighted with a red circle.

NAME	DEFINITION	REF.	LIFECYCLE	SYSTEM SHORT NAME	ATTRIBUTES	DATA QUALITY
CUSTOMER	This is an object created for CUST...	DS-1	Draft	S1	5 attributes	99%
ORDERS	This is an object created for ORDE...	DS-4	Draft	S4	4 attributes	96%



## Informatica Data Marketplace – Consumers Must Agree To The Terms And Conditions When Submitting An Order

Copyright © Intelligent Business Strategies 1992-2024

131



## Informatica Axon Data Marketplace – Fulfill And Track All Sharing Activity Includes Dialogue With Data Owner To Gain Approval To Access The Data

- Deliver and auto-provision data after approval of request
- Monitor and audit delivery at the order level
- Integrate order processing with data engineering processes (JIRA, ServiceNow, custom)

132





## Data Marketplace Example Amazon DataZone

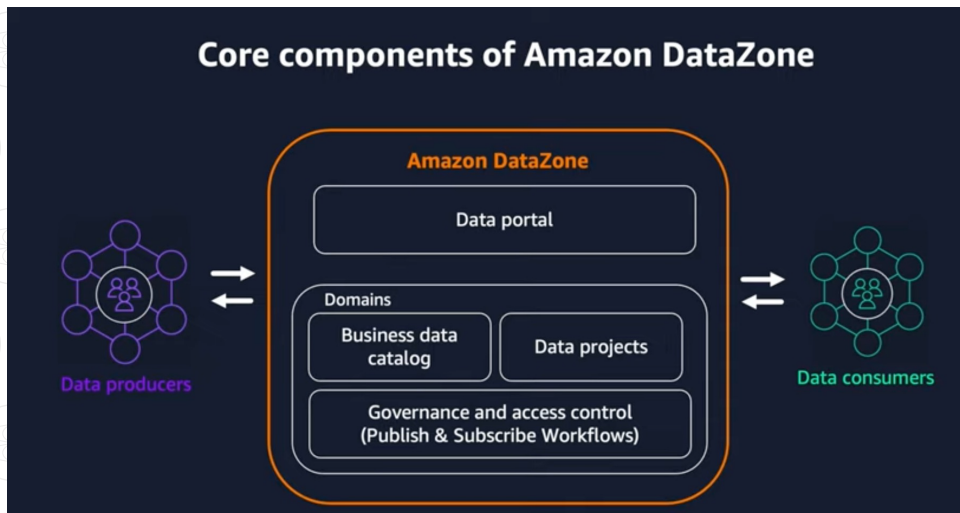


Image source & copyright: AWS

Copyright © Intelligent Business Strategies 1992-2024

133



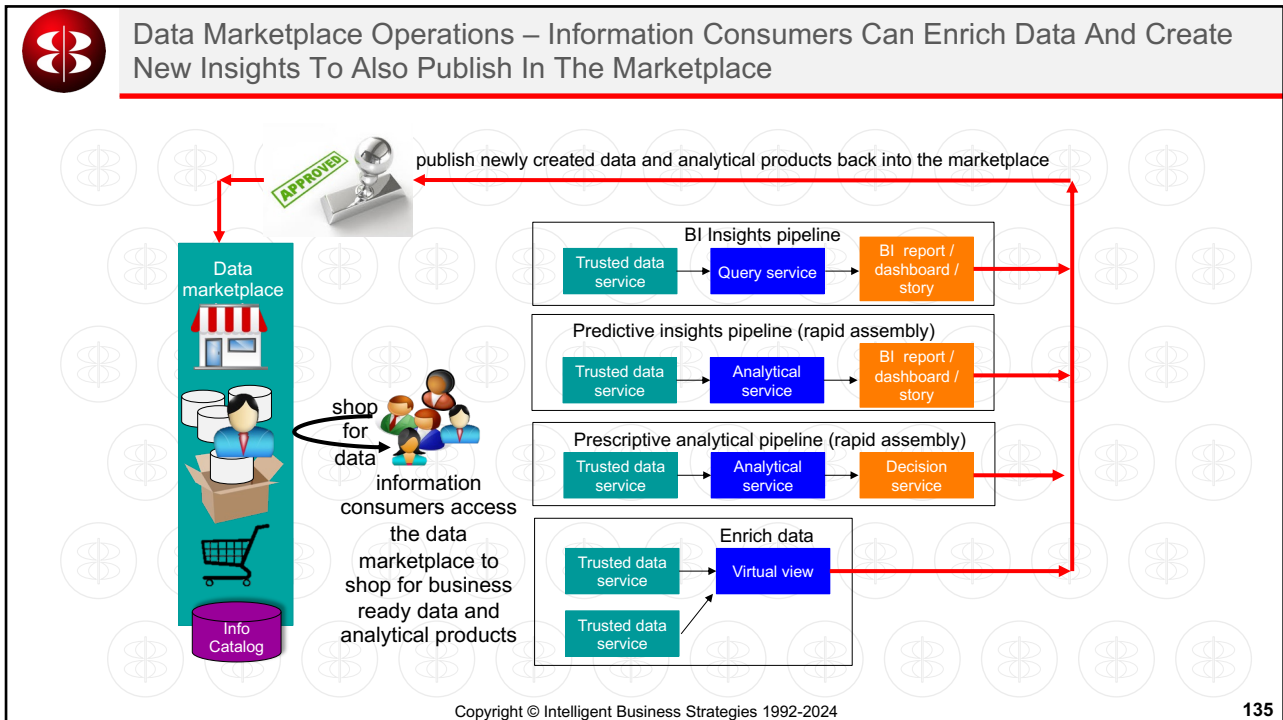
## Amazon DataZone – Publish Data With Publishing Agreement

The screenshot shows the Amazon DataZone user interface. The main heading is "Publish data". Below it, there are two options for publishing: "Automated publish" (selected) and "Manual publish". The "Automated publish" option includes a "Publishing job details" section with fields for "Name" and "Description", and a "Publishing agreement" section where "Sales Producer Project default agreement" is selected. To the right, there is a "Sales Producer Project" overview card. This card has tabs for "SUBSCRIBED DATA", "PUBLISHED DATA", "PENDING", and "SETTINGS". The "PUBLISHED DATA" tab is active, showing a table with one entry: "catalog\_sales" (type: TABLE, size: 100 MB, last updated: Mar 23, 2023, 10:55 AM PST). A red box highlights this entry. Below the table, it says "Published by sales-publishing-job Publishing agreement: Sales Producer Project default agreement". On the far right, there are "PROJECT RESOURCES" and "ANALYTICAL TOOLS" sections.

Source: AWS

Copyright © Intelligent Business Strategies 1992-2024

134

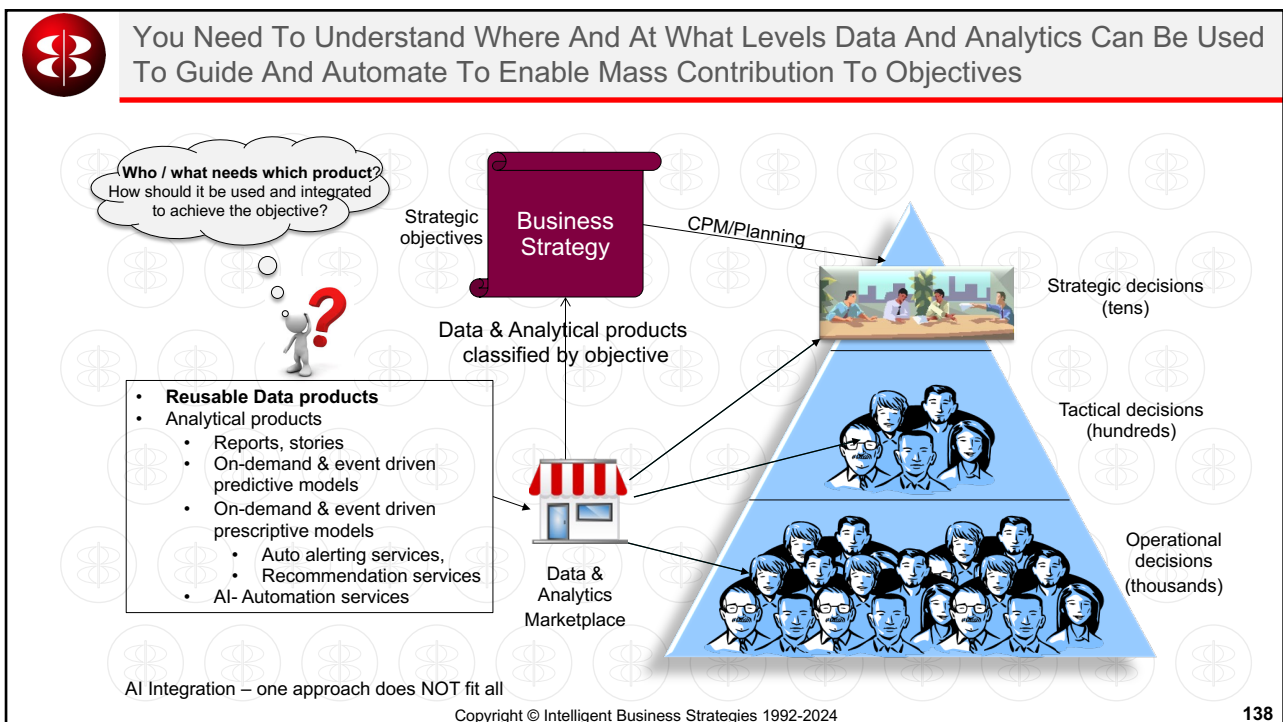
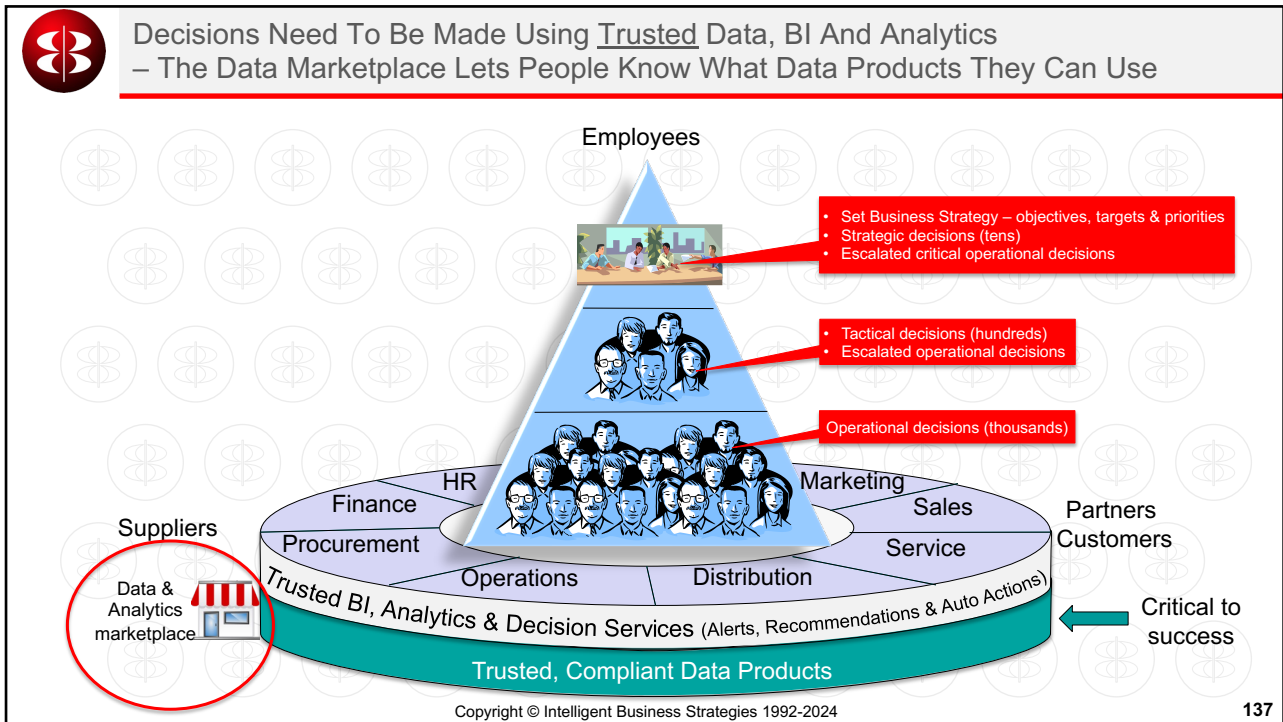


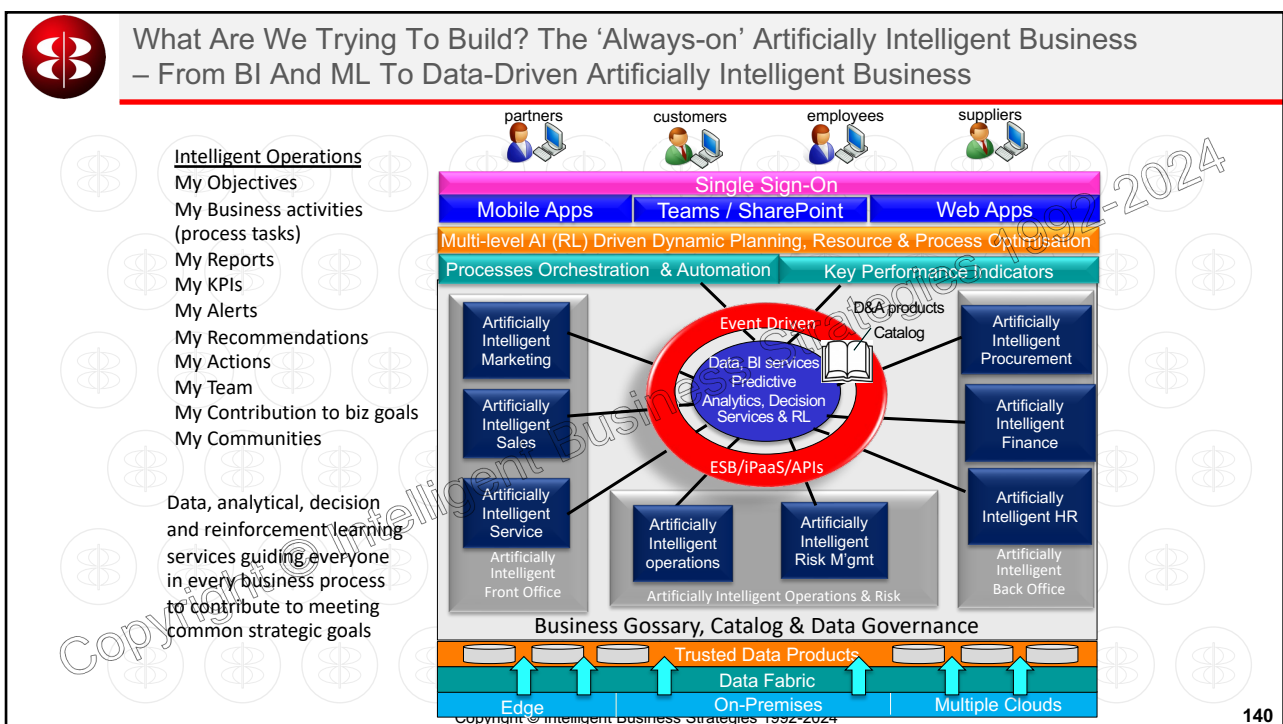
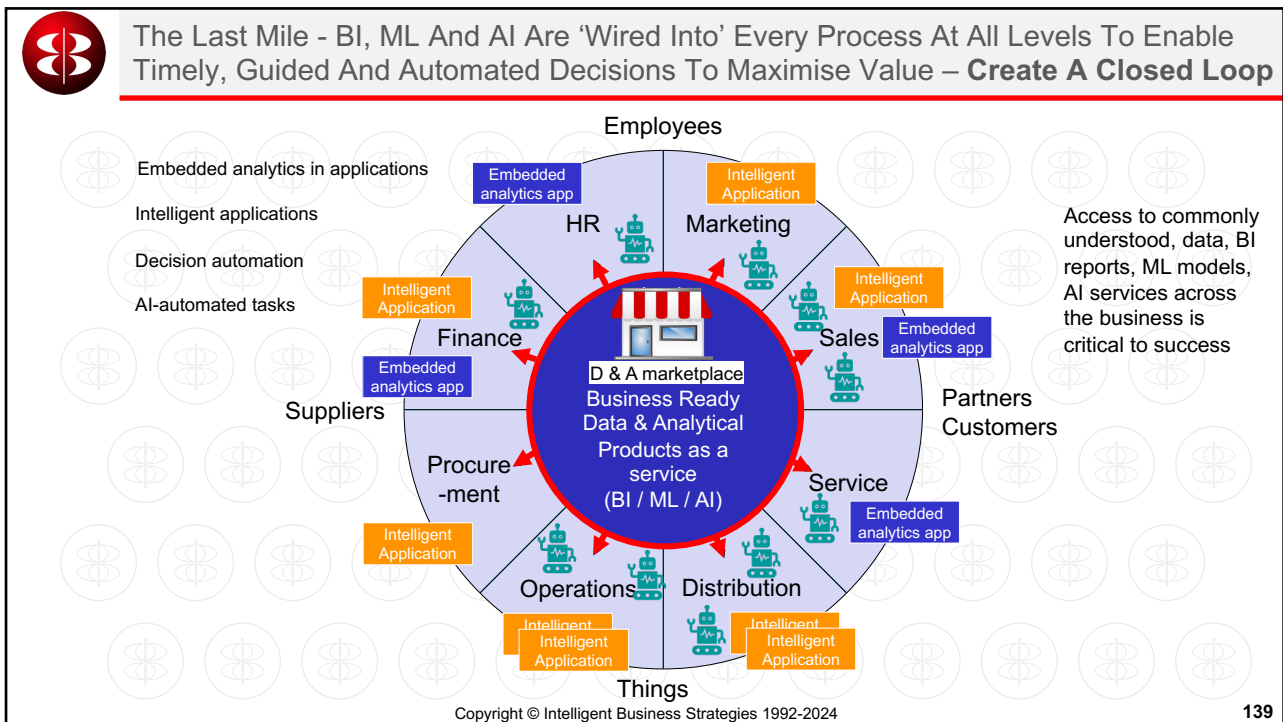
### Key Elements In Enterprise Data Marketplace Operations – Monitoring

- Who uses this data product
- Data product usage frequency
  - How many people and applications make use of a data product
  - Identify unused data products
- Navigational patterns
  - Most popular paths to finding and consuming a data product
  - Data products looked at but not consumed
  - Most popular paths leading to consumption abandoned
- Determine experts
  - Most frequent users of data products associated with a particular entity or business area
- Self-learn (AI)

Copyright © Intelligent Business Strategies 1992-2024

136







## About Mike Ferguson



[www.intelligentbusiness.biz](http://www.intelligentbusiness.biz)



[mferguson@intelligentbusiness.biz](mailto:mferguson@intelligentbusiness.biz)



@mikeferguson1



(+44) 1625 520700

Mike Ferguson is Managing Director of Intelligent Business Strategies Limited. As an independent IT industry analyst and consultant, he specialises in BI / analytics and data management. With over 40 years of IT experience, Mike has consulted for dozens of companies on BI/Analytics, data strategy, technology selection, enterprise architecture, and data management. Mike is also conference chairman of Big Data LDN, the largest data and analytics conference in Europe and a member of the EDM Council CDMC Executive Advisory Board. He has spoken at events all over the world and written numerous articles. Formerly he was a principal and co-founder of Codd and Date – the inventors of the Relational Model that caused the birth of relational databases and SQL, Chief Architect at Teradata on the Teradata DBMS and European Managing Director of Database Associates. He teaches popular master classes in Data Strategy, Data Catalogs, Data Warehouse Modernisation, Practical Guidelines for Implementing a Data Mesh, Big Data Fundamentals, How to Govern Data Across a Distributed Data Landscape, Machine Learning and Advanced Analytics, and Embedded Analytics, Intelligent Apps and AI Automation



Thank You!

Copyright © Intelligent Business Strategies 1992-2024

141